ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA: 3.64) in the Third Cycle and Graded as Category I University by MHRD- UGC]
(A State University Established by the Government of Tamilnadu)

KARAIKUDI-630 003

DIRECTORATE OF DISTANCE EDUCATION

M.Sc. (MICROBIOLOGY)

36443–BIOINFORMATICS & BIOSTATISTICS

SYLLABI-BOOK MAPPING TABLE AND INSTRUCTION

CONTENTS

BLOCK 1: Introduction to Bioinformatics

UNIT-I

Structure

NOTES

1.1 Introduction

Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modeling of data for analysis, visualization or prediction through the development of algorithms and software. In this unit describes the computational application of biology and needs. Thus, by studying the computers, it is very essential to know the relationship between computer with biology to the gain knowledge like servers, workstation, data storage, data retrieval, usage and characteristic of computers. Hence, it is essential to one who is willing to access the computers and servers.

1.2 Objectives

After going through the unit you will able to;

- Understand the concept of computer age
- Relate the meaning of computers with biology
- Arise the computational approaches to biological questions

NOTES

- Identify the importance and key role of servers and workstation.

1.3 Biology in the computer age

Biology is defined as the study of living thing to analyze the interaction of species and populations, to the function of tissues and cells within an individual organism. The computers in the biology which arise when the data are collected, stored, and interpret now, at the beginning of the 21^{st} century, we use sophisticated laboratory technology that allows us to collect data faster than we can interpret it. We have vast volumes of DNA sequence data at our fingertips. But how do we figure out which parts of that DNA control the various chemical processes of life? We know the function and structure of some proteins, but how do we determine the function of new proteins? And how do we predict what a protein will look like, based on knowledge of its sequence? We understand the relatively simple code that translates DNA into protein. But how do we find meaningful new words in the code and add them to the DNA-protein dictionary? To overcome all these issues, Bioinformatics, a new emerging field arises to answer all the questions.

Bioinformatics is the science of using information to understand biology; it's the tool we can use to help us answer these questions and many others like them.

NOTES

Unfortunately, with all the hype about mapping the human genome, bioinformatics has achieved buzzword status; the term is being used in a number of ways, depending on who is using it. Strictly speaking, bioinformatics is a subset of the larger field of *computational biology,* the application of quantitative analytical techniques in modeling biological systems. In this book, we stray from bioinformatics into computational biology and back again. The distinctions between the two aren't important for our purpose here, which is to cover a range of tools and techniques we believe are critical for molecular biologists who want to understand and apply the basic computational tools that are available today. The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition. Researchers come to bioinformatics from many fields, including mathematics, computer science, and linguistics. Unfortunately, biology is a science of the specific as well as the general. Bioinformatics is full of pitfalls for those who look for patterns and make predictions without a complete understanding of where biological data comes from and what it means. By providing algorithms, databases, user interfaces, and statistical tools, bioinformatics makes it possible to do exciting things such as compare DNA sequences and generate results that are potentially

NOTES

significant. "Potentially significant" is perhaps the most important phrase. These new tools also give you the opportunity to over-interpret data and assign meaning where none really exists. We can't overstate the importance of understanding the limitations of these tools. But once you gain that understanding and become an intelligent consumer of bioinformatics methods, the speed at which your research progresses can be truly amazing.

1.3.1 How Is Computing Changing Biology?

An organism's hereditary and functional information is stored as DNA, RNA, and proteins, all of which are linear chains composed of smaller molecules. These macromolecules are assembled from a fixed alphabet of well-understood chemicals: DNA is made up of four deoxyribonucleotides (adenine, thymine, cytosine, and guanine), RNA is made up from the four ribonucleotides (adenine, uracil, cytosine, and guanine), and proteins are made from the 20 amino acids. Because these macromolecules are linear chains of defined components, they can be represented as sequences of symbols. These sequences can then be compared to find similarities that suggest the molecules are related by form or function.

Sequence comparison is possibly the most useful computational tool to emerge for molecular biologists.

NOTES

The World Wide Web has made it possible for a single public database of genome sequence data to provide services through a uniform interface to a worldwide community of users. With a commonly used computer program called fsBLAST, a molecular biologist can compare an uncharacterized DNA sequence to the entire publicly held collection of DNA sequences. In the next section, we present an example of how sequence comparison using the BLAST program can help you gain insight into a real disease.

1.3.2 The Eye of the Fly

Fruit flies (*Drosophila melanogaster*) are a popular model system for the study of development of animals from embryo to adult. Fruit flies have a gene called *eyeless*, which, if it's "knocked out" (i.e., eliminated from the genome using molecular biology methods), results in fruit flies with no eyes. It's obvious that the *eyeless* gene plays a role in eye development.

Researchers have identified a human gene responsible for a condition called *aniridia*. In humans who are missing this gene (or in whom the gene has mutated just enough for its protein product to stop functioning properly), the eyes develop without irises. If the gene for *aniridia* is inserted into an eyeless drosophila "knock out," it causes the production of normal drosophila eyes. It's an interesting coincidence. Could

NOTES

there be some similarity in how *eyeless* and *aniridia* function, even though flies and humans are vastly different organisms? To gain insight into how *eyeless* and *aniridia* work together, we can compare their sequences. Always bear in mind, however, that genes have complex effects on one another. Careful experimentation is required to get a more definitive answer.

As little as 15 years ago, looking for similarities between *eyeless* and *aniridia* DNA sequences would have been like looking for a needle in a haystack. Most scientists compared the respective gene sequences by hand-aligning them one under the other in a word processor and looking for matches character by character. This was time- consuming, not to mention hard on the eyes.

In the late 1980s, fast computer programs for comparing sequences changed molecular biology forever. Pairwise comparison of biological sequences is the foundation of most widely used bioinformatics techniques. Many tools that are widely available to the biology community-including everything from multiple alignment, phylogenetic analysis, motif identification, and homology-modeling software, to web- based database search services-rely on pairwise sequence-comparison algorithms as a core element of their function.

NOTES

1.3.4 The First Information Age in Biology

Biologists have been dealing with problems of information management since the 17th century.

The roots of the concept of evolution lie in the work of early biologists who catalogued and compared species of living things. The cataloguing of species was the preoccupation of biologists for nearly three centuries, beginning with animals and plants and continuing with microscopic life upon the invention of the compound microscope. New forms of life and fossils of previously unknown, extinct life forms are still being discovered even today.

All this cataloguing of plants and animals resulted in what seemed a vast amount of information at the time. In the mid-16th century, Otto Brunfels published the first major modern work describing plant species, the *Herbarium vitae eicones*. As Europeans traveled more widely around the world, the number of catalogued species increased, and botanical gardens and herbaria were established. The number of catalogued plant types was 500 at the time of Theophrastus, a student of Aristotle. By 1623, Casper Bauhin had observed 6,000 types of plants. Not long after John Ray introduced the concept of distinct species of animals and plants, and developed guidelines based on anatomical features for distinguishing conclusively between species. In the

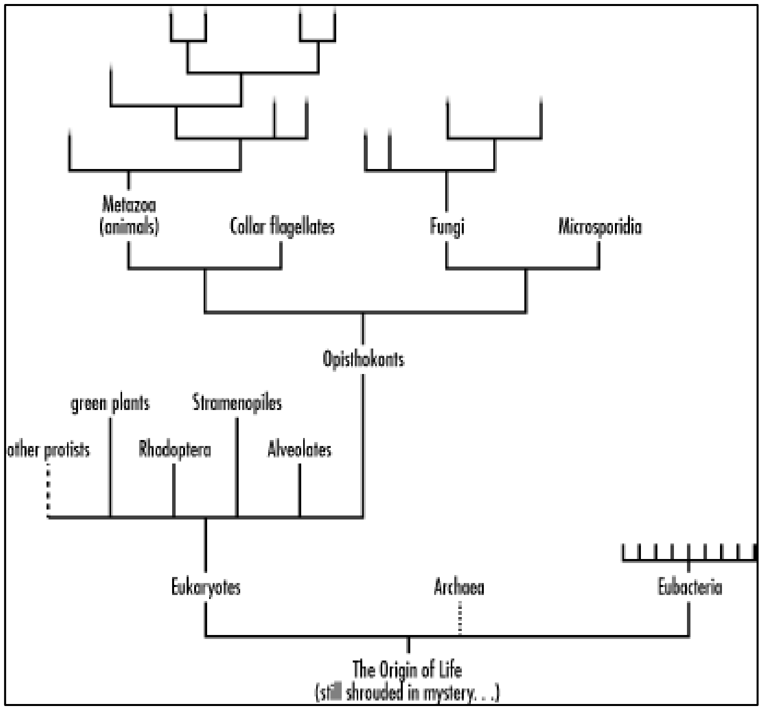
1730s, Carolus Linnaeus catalogued 18,000 plant species and over 4,000 species of animals, and established the basis for the modern taxonomic naming system of kingdoms, classes, genera, and species. By the end of the 18th century, Baron Cuvier had listed over 50,000 species of plants.

It was no coincidence that a concurrent preoccupation of biologists, at this time of exploration and cataloguing, was classification of species into an orderly taxonomy. A botany text might encompass several volumes of data, in the form of painstaking illustrations and descriptions of each species encountered. Biologists were faced with the problem of how to organize, access, and sensibly add to this information. It was apparent to the casual observer that some living things were more closely related than others. A rat and a mouse were clearly more similar to each other than a mouse and a dog. But how would a biologist know that a rat was like a mouse (but that rat was not just another name for mouse) without carrying around his several volumes of drawings? A nomenclature that uniquely identified each living thing and summed up its presumed relationship with other living things, all in a few words, needed to be invented.

The solution was relatively simple, but at the time, a great innovation. Species were to be named with a

NOTES

NOTES



series of one-word names of increasing specificity. First a very general division was specified: animal or plant? This was the kingdom to which the organism belonged. Then, with increasing specificity, came the names for class, genera, and species. This schematic way of classifying species, as illustrated in

NOTES

A modern taxonomy of the earth's millions of species is too complicated for even the most zealous biologist to memorize, and fortunately computers now provide a way to maintain and access the taxonomy of species. The University of Arizona's Tree of Life project and NCBI's Taxonomy database are two examples of online taxonomy projects.

Taxonomy was the first informatics problem in biology. Now, biologists have reached a similar point of information overload by collecting and cataloguing information about individual genes. The problem of organizing this information and sharing knowledge with the scientific community at the gene level isn't being tackled by developing a nomenclature. It's being attacked directly with computers and databases from the start.

The evolution of computers over the last half-century has fortuitously paralleled the developments in the physical sciences that allow us to see biological systems in increasingly fine detail. Figure 1.2 illustrates the astonishing rate at which biological knowledge has expanded in the last 20 years.

NOTES

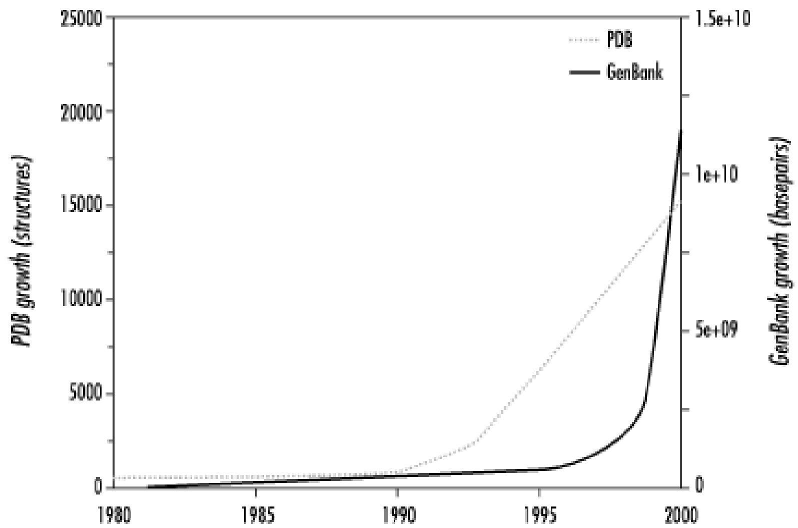


Figure 1.2 The growth of GenBank and the PDB

NOTES

1.4 Computational Approaches to Biological Questions

There is a standard range of techniques that are taught in bioinformatics courses. Currently, most of the important techniques are based on one key principle: that sequence and structural homology (or similarity) between molecules can be used to infer structural and functional similarity. In this unit, we will give you an overview of the standard computer techniques available to biologists and we will discuss how specific software packages implement these techniques and how you should use them.

1.4.1 Molecular Biology's Central Dogma

Before we go any further, it's essential that you understand some basics of cell and molecular biology. If you're already familiar with DNA and protein structure, genes, and the processes of transcription and translation, feel free to skip ahead to the next section.

The central dogma of molecular biology states that:

DNA acts as a template to replicate itself,

DNA is also transcribed into RNA, and

RNA is translated into protein.

As you can see, the central dogma sums up the function of the genome in terms of information. Genetic information is conserved and passed on to progeny through the process of replication. Genetic information is

NOTES

also used by the individual organism through the processes of transcription and translation. There are many layers of function, at the structural, biochemical, and cellular levels, built on top of genomic information. But in the end, all of life's functions come back to the information content of the genome.

Put another way, genomic DNA contains the master plan for a living thing. Without DNA, organisms wouldn't be able to replicate themselves. The raw "one-dimensional" sequence of DNA, however, doesn't actually do anything biochemically; it's only information, a blueprint if you will, that's read by the cell's protein synthesizing machinery. DNA sequences are the punch cards; cells are the computers.

DNA is a linear polymer made up of individual chemical units called *nucleotides* or *bases*. The four nucleotides that make up the DNA sequences of living things (on Earth, at least) are adenine, guanine, cytosine, and thymine—designated A, G, C, and T, respectively. The order of the nucleotides in the linear DNA sequence contains the instructions that build an organism. Those instructions are read in processes called replication, transcription, and translation.

1.4.2 Replication of DNA

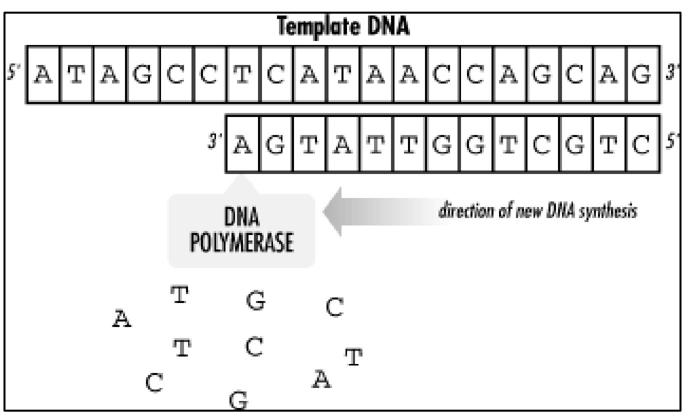
The unusual structure of DNA molecules gives DNA special properties. These properties allow the

NOTES

information stored in DNA to be preserved and passed from one cell to another, and thus from parents to their offspring. Two molecules of DNA form a double-helical structure, twining around each other in a regular pattern along their full length—which can be millions of nucleotides. The halves of the double helix are held together by bonds between the nucleotides on each strand. The nucleotides also bond in particular ways: A can pair only with T, and G can pair only with C. Each of these pairs is referred to as a *base pair*, and the length of a DNA sequence is often described in base pairs (or bp), kilobases (1,000 bp), megabases (1 million bp), etc.

Each strand in the DNA double helix is a chemical "mirror image" of the other. If there is an A on one strand, there will always be a T opposite it on the other. If there is a C on one strand, its partner will always be a G.

When a cell divides to form two new *daughter cells*, DNA is *replicated* by untwisting the two strands of the double helix and using each strand as a template to build its chemical mirror image, or *complementary strand*. This process is illustrated in Figure 1.3.



NOTES

1.4.3 Genomes and Genes

The entire DNA sequence that codes for a living thing is called its *genome*. The genome doesn't function as one long sequence, however. It's divided into individual genes. A *gene* is a small, defined section of the entire genomic sequence, and each gene has a specific, unique purpose.

There are three classes of genes. *Protein-coding* genes are templates for generating molecules called proteins. Each *protein* encoded by the genome is a chemical machine with a distinct purpose in the organism. *RNA-specifying* genes are also templates for chemical machines, but the building blocks of RNA machines are different from those that make up proteins. Finally, *untranscribed genes* are regions of genomic DNA that have some functional purpose but don't achieve that purpose by being transcribed or translated to create another molecule.

1.4.4 Transcription of DNA

DNA can act not only as a template for making copies of itself but also as a blueprint for a molecule called ribonucleic acid (RNA). The process by which DNA is transcribed into RNA is called *transcription* and is illustrated in Figure 1.4. RNA is structurally similar to DNA. It's a polymeric molecule made up of individual chemical units, but the chemical backbone that holds these

NOTES

units together is slightly different from the backbone of DNA, allowing RNA to exist in a single-stranded form as well as in a double helix. These single-stranded molecules still form base pairs between different parts of the chain, causing RNA to fold into 3D structures. The individual chemical units of RNA are designated A, C, G, and U (uracil, which takes the place of thymine).

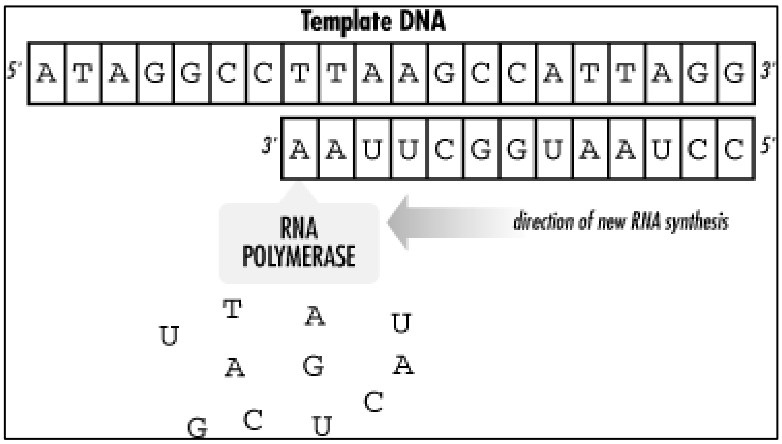


Figure 1.4. Schematic of DNA being transcribed into RNA

The genome provides a template for the synthesis of a variety of RNA molecules: the three main types of RNA are messenger RNA, transfer RNA, and ribosomal RNA. *Messenger* RNA (mRNA) molecules are RNA transcripts of genes. They carry information from the genome to the ribosome, the cell's protein synthesis apparatus. *Transfer* RNA (tRNA) molecules are untranslated RNA molecules

NOTES

that transport amino acids, the building blocks of proteins, to the ribosome. Finally, *ribosomal* RNA (rRNA) molecules are the untranslated RNA components of ribosomes, which are complexes of protein and RNA. rRNAs are involved in anchoring the mRNA molecule and catalyzing some steps in the translation process. Some viruses also use RNA instead of DNA as their genetic material.

1.4.5 Translation of mRNA

Translation of mRNA into protein is the final major step in putting the information in the genome to work in the cell. Like DNA, proteins are linear polymers built from an alphabet of chemically variable units. The protein alphabet is a set of small molecules called *amino acids*.

Unlike DNA, the chemical sequence of a protein has physicochemical "contvent" as well as information content. Each of the 20 amino acids commonly found in proteins has a different chemical nature, determined by its side chain—a chemical group that varies from amino acid to amino acid. The chemical sequence of the protein is called its *primary structure*, but the way the sequence folds up to form a compact molecule is as important to the function of the protein as is its primary structure. The secondary and tertiary structure elements that make up the protein's final fold can bring distant parts of the

chemical sequence of the protein together to form functional sites.

As shown in Figure 1.5, the *genetic code* is the code that translates DNA into protein. It takes three bases of DNA (called a *codon*) to code for each amino acid in a protein sequence. Simple combinatorics tells us that there are 64 ways to choose 3 nucleotides from a set of 4, so there are 64 possible codons and only 20 amino acids. Some codons are redundant; others have the special function of telling the cell's translation machinery to stop translating an mRNA molecule. Figure 1.6 shows how RNA is translated into protein.

NOTES

NOTES

First Position		Second Position: U		Second Position: C		Second Position: A		Second Position: G		Third Position
	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
		UUC		UCC		UAC		UGC		C
		UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
		UUG		UCG		UAG	Stop	UGG	Trp	G
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
		CUC		CCC		CAC		CGC		C
		CUA		CCA		CAA	Gln	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	Lys	AGA	Arg	A
		AUG	Met (start)	ACG		AAG		AGG		G
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	Glu	GGA		A
		GUG		GCG		GAG		GGG		G

Figure 1.5. The genetic code

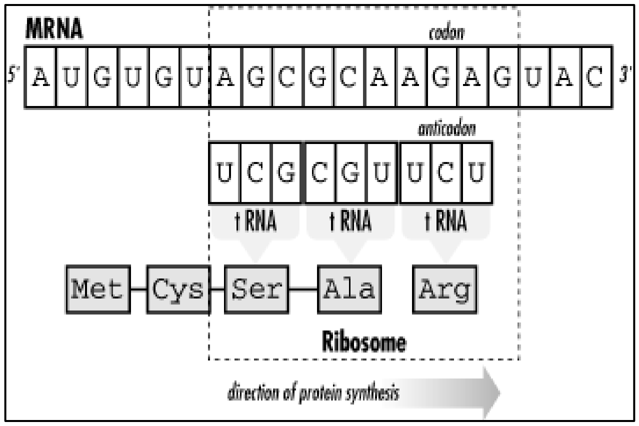


Figure 1.6. Synthesis of protein with standard base pairing

NOTES

1.4.6 Molecular Evolution

Errors in replication and transcription of DNA are relatively common. If these errors occur in the reproductive cells of an organism, they can be passed to its progeny. Alterations in the sequence of DNA are known as *mutations*. Mutations can have harmful results—results that make the progeny less likely to survive to adulthood. They can also have beneficial results, or they can be neutral. If a mutation doesn't kill the organism before it reproduces, the mutation can become fixed in the population over many generations. The slow accumulation of such changes is responsible for the process known as *evolution*. Access to DNA sequences gives us access to a more precise understanding of evolution. Our understanding of the molecular mechanism of evolution as

NOTES

a gradual process of accumulating DNA sequence mutations is the justification for developing hypotheses based on DNA and protein sequence comparison.

1.5 Basics of computers

A computer is a programmable machine designed to perform arithmetic and logical operations automatically and sequentially on the input given by the user and gives the desired output after processing. Computer components are divided into two major categories namely hardware and software. Hardware is the machine itself and its connected devices such as monitor, keyboard, mouse etc. Software are the set of programs that make use of hardware for performing various functions.

1.5.1 CHARACTERISTICS OF COMPUTERS

Speed

Computers work at an incredible speed. A powerful computer is capable of performing about 3-4 million simple instructions per second.

Accuracy

In addition to being fast, computers are also accurate. Errors that may occur can almost always be attributed to human error (inaccurate data, poorly designed system or faulty instructions/programs written by the programmer)

Diligence

Unlike human beings, computers are highly consistent. They do not suffer from human traits of boredom and tiredness resulting in lack of concentration. Computers, therefore, are
better than human beings in performing voluminous and repetitive jobs.

Versatility

Computers are versatile machines and are capable of performing any task as long as it can be broken down into a series of logical steps. The presence of computers can be seen in almost every sphere – Railway/Air reservation, Banks, Hotels, Weather forecasting and many more.

Storage Capacity

Today's computers can store large volumes of data. A piece of information once recorded (or stored) in the computer, cannot never be forgotten and can be retrieved almost instantaneously.

1.6 Servers and workstation

A complete understanding of server hardware can take a great deal of study. This chapter provides an overview of the server and identifies different server types and their roles. You will be introduced to some of the hardware that makes the server unique from the ordinary PC, and you will learn about RAID and other storage systems.

1.6.1 Server Types and Services

`Servers provide a variety of services. Some of the services a server can provide are authentication and security, Web, mail, and print. A server can be called by many names. For example, it

NOTES

can be called an authentication and security database server, Web server, mail server, and print server, Figure 9-1. A network may have a single server that provides a variety of services, or it may have a group of servers, each providing a specifi c service. A small network usually has one server set up to handle many different services. A large network usually has several servers, each providing a different service or set of services. For example, a large corporation may use one server to handle e-mail requests and Web hosting, another server to serve as a domain controller to provide security for the entire network, and another server to provide application software, a database for its clients, and support for print operations. The more services a server provides and the more clients it services creates a higher demand on the server. A single server with a limited amount of services may be fine for a relatively small number of users, typically less than 25. A commercial enterprise spanning across the country or world may require hundreds of servers. The network administrator or network designer decides the number and capacity of individual servers needed by taking into account the number of users and the systems predicted network traffic. Each network system is uniquely designed, even though each network has many similarities. Some network equipment providers have software programs that help you design a network. You simply enter information, such as the number of clients, offices, cities, and countries and the type of software and services to be provided. After all the information is

collected, the software program provides an estimate of the size and number of servers required.

1.6.2 Thin Servers

A thin server is a server that has only the hardware and software needed to support and run a specifi c function, such as Web services, print services, and file services. It is more economical to use a thin server as a print server than to tie up a more expensive server simply to handle printing on a network. IBM markets a thin server, which consists of a sealed box that contains only the essential hardware and software required for supporting the server's dedicated function.

1.6.3 Thin Client Servers

A thin client server is a server that provides applications and processing power to a thin client. Thin client servers run terminal server software and may have more RAM and hard disk drive storage than needed. A thin client is a computer that typically has a minimal amount of processing power and memory. It relies on the server's processor and memory for processing data and running software applications. Thin clients are becoming common in industry for such applications as hotel bookings, airline ticketing, and medical record access. When budgets are tight, you can install what is normally considered an obsolete PC on a thin client network. The obsolete PC is obsolete because of its processing speed, lack of storage, and inability to run new software. The thin client server can provide all the services that are required by the obsolete PC, thus making the obsolete PC a useful

NOTES

workstation. Do not confuse a thin client with a dumb terminal. A dumb terminal sends user input to a mainframe. Dumb terminals have absolutely no computing power, operating system, hard disk drive, BIOS, and RAM. A thin client may not need a hard disk drive or an operating system; however, it is still a full-fledged computer because it has a CPU and processing power. Windows 2000 Server, Windows Server 2003, and Windows Server 2008 come with Terminal Services software. Once Terminal Services is set up on the server, you have the option to generate a disk that will assist you in automatically setting up thin client workstations.

1.7 LET US SUM UP

In this unit, you have learnt about the computer age, computational approaches to biological questions and characteristics of computers. This knowledge would make understand what is computer and how it plays major role biology. The concept such as computer age has given detail knowledge about how computers interpret and evolution of biology. The application of computer age would pave the key attention to the student. The servers and workstation gave a how the networks are interlinked and the main access worldwide.

1.8 Unit-End Exercise

What are the main characteristics of computers?

What is bioinformatics?

Write a role on bioinformatics in modern world?

NOTES

1.9 CHECK YOUR PROGRESS

1. Comment on computational approaches to biological question.
2. What is Bioinformatics?
3. Give a note on different types of servers

1.10 ANSWERS TO CHECK YOUR PROGRESS

1**.** Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modeling of data for analysis, visualization or prediction through the development of algorithms and software.

2. Biological software's information, Protein structure identification, structural and functional annotation, how to retrieve the data and etc.

3. Speed, Accuracy, Diligence, Versatility and Storage Capacity

4. Thin and Thin Client servers.

1.11 SUGGESTED READINGS

1. Cynthia Gibas, Per Jambeck (2001), Developing Bioinformatics Computer Skills, O'Reilly Media publication

2. Sherman (2001), Basic Computer Skills made easy, Butterworth-Heinemann Ltd, USA.

3. Balaguruswamy E (1985), Computer Fundamentals and Applications (2nd Ed.), Tata McGraw-Hill Publishing Co. Ltd., India.

UNIT-II

NOTES

Structure

NOTES

2.1 Introduction

An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer. In this unit, we cover the software's related information to your study. The operating system which gives details knowledge about the computer progress. The basic characteristics and working mechanism of UNIX and LINUX will be studied in detail. Hence the unit would pave the attention to the student for working the computers like access operating system, search engine world wide web and finding and collecting article through their study.

2.2 Objectives

After going through the unit you will able to;

- Understand the concept of operating system

NOTES

- Know about UNIX and Linux operating system and their evolution.
- To identify the scientific articles and their related databases
- The data encrypting and detail difference network analysis will be studied

2.3 Operating systems

An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer. There are many OS were available in the market such as Linux, Windows, Mac OS, Solaris and etc. The above mentioned OS is freely and commercially available. In our unit we cover UNIX and Linux OS. These two OS is freely available and graphical user interface (GUI) system.

2.3.1 UNIX OPERATING SYSTEM

UNIX is a powerful operating system for multiuser computer systems. It has been in existence for over 25 years, and during that time has been used primarily in industry and academia, where networked systems and multiuser high-performance computer systems are required. UNIX is optimized for tasks that are only fairly recent additions to personal-computer operating systems,

NOTES

or which are still not even available in some PC operating systems: networking with other computers, initiating multiple asynchronous tasks, retaining unique information about the work environments of multiple users, and protecting the information stored by individual users from other users of the system. Unix is the operating system of the World Wide Web; the software that powers the Web was invented in Unix, and many if not most web servers run on Unix servers.

UNIX is rich in commands and possibilities. Every distribution of UNIX comes with a powerful set of built-in programs. Everything from networking software to word-processing software to electronic mail and news readers is already a part of UNIX. Many other programs can be downloaded and installed on UNIX systems for free.

It might seem that there's far too much to learn to make working on a UNIX system practical. It's possible, however, to learn a subset of UNIX and to become a productive UNIX user without knowing or using every program and feature.

2.3.2 UNIX Shell

When you log into a UNIX system or open a new window in your system's window manager interface, the system automatically starts a program called a shell for you. The shell program interprets the commands you enter and

provides you with a working environment and an interface to the operating system. It's possible to work in UNIX without the shell using graphical file manager tools, but you'll find that many shell commands are useful for data processing and analysis.

2.3.3 Standard Input and Output

By default, many UNIX commands read from standard input and send their output to standard output. Standard input and output are file descriptors associated with your terminal. A program reading from standard input will simply hang out and wait for you to type something on your keyboard and press the Enter key. A program writing to standard output spews its output to your terminal, sometimes far faster than you can read it.

Some UNIX commands read a hyphen (-) surrounded by whitespace on either side as "data from standard input." This construct can then be used in place of a filename in the command line. Absence of an output filename is sufficient to cause the program to write to standard output.

2.3.4 Graphical Interfaces for UNIX

Although UNIX is a text-based operating system, you no longer have to experience it as a black screen full of glowing green or amber letters. Most UNIX systems use a variant of the X Window System. The X Window System formats the screen environment and allows you to have

NOTES

multiple windows and applications open simultaneously. X windows are customizable so that you can use menu bars and other widgets much like PC operating systems. Individual UNIX shells on the host machine as well as on networked machines are opened as windows, allowing you to exploit UNIX's multitasking capabilities and to have many shells active simultaneously. In addition to UNIX shells and tools, there are many applications that take advantage of the X system and use X windows as part of their graphical user interfaces, allowing these applications to be run while still giving access to the UNIX command line.

The GNOME and KDE desktop environments, which are included in most major Linux distributions, make your Linux system look even more like a personal computer. Toolbars, visual file managers, and a configurable desktop replicate the feeling of a Windows or Mac work environment, except that you can also open a shell window and run UNIX programs.

2.3.5 Linux

Linux (LIH-nucks) is an open source version of UNIX, named for its original developer, Linus Torvalds of the University of Helsinki in Finland. Originally undertaken as a one-man project to create a free UNIX for personal computers, Linux has grown from a hobbyist

NOTES

project into a product that, for the first time, gives the average personal-computer user access to a UNIX system.

In this unit, we focus on Linux for three reasons. First, with the availability of Linux, UNIX is cheap (or free, if you have the patience to download and install it). Second, under Linux, inexpensive PCs regarded as "obsolete" by Windows users become startlingly flexible and useful workstations. The Linux operating system can be configured to use a much smaller amount of system resources than the personal computer operating systems, which means computers that have been outgrown by the ever-expanding system requirements of PC programs and operating systems, can be given a new lease on life by being reconfigured to run Linux. Third, Linux is an excellent platform for developing software, so there's a rich library of tools available for computational biology and for research in general.

2.3.6 Installing the software in Linux

1. Download the source code distribution. Use binary mode; compressed text files are encoded.
2. Download and read the instructions (usually a README or INSTALL file; sometimes you have to find it after you extract the archive).
3. Make a new directory and move the archive into it, if necessary.
4. Uncompress (*.Z) or gunzip (*.gz) the file.

NOTES

5. Extract the archive using tar xvf or as instructed.
6. Follow the instructions (did we say that already?).
7. Run the configuration script, if present.
8. Run make if a Makefile is present.
9. If a Makefile isn't present and all you see are *.f or *.c files, use gcc or g77 to compile them, as discussed earlier.
10. Run the installation script, if present.
11. Link the newly created binary executable into one of the binary-containing directories in your path using ln -s (this is usually part of the previous step, but if there is no installation script, you may need to create the link by hand).

2.3.7 Paths to Files and Directories

Each file on the filesystem can be uniquely identified by a combination of a filename and a path. You can reference any file on the system by giving its full name, which begins with a / indicating the root directory, continues through a list of subdirectories (the components of the path) and ends with the filename. The full name, or *absolute path,* of a file in someone's home directory might look like this:/home/jambeck/mustelidae/weasels.txt

2.3.8 Applications of Linux

Linux is a free available OS for deskotop and laptop platform with GUI constructing interface. Many popular applications such as

Mozilla Firefox, OpenOffice.org/LibreOffice and Blender has available in Linux OS. Besides externally visible components, such as X window managers, a non-obvious but quite central role is played by the programs hosted by freedesktop.org, such as D-Bus or PulseAudio; both major desktop environments (GNOME and KDE) include them, each offering graphical front-ends written using the corresponding toolkit (GTK+ or Qt). A display server is another component, which for the longest time has been communicating in the X11 display server protocol with its clients; prominent software talking X11 includes the X.Org Server and Xlib. Linux distributions have long been used as server operating systems

2.4 World-Wide Web

The World-Wide Web is a collection of documents and services, distributed across the Internet and linked together by hypertext links. The web is therefore a subset of the Internet, not the same thing. Much of what is available on the web consists of web documents, which are often (somewhat misleadingly) called “home pages.”

A home page appears to be a single entity, but is actually made up of a number of separate and distinct files, which may incorporate one or more of the following, in different configurations: text

*graphics

*animation

NOTES

*audio

*video

*hyperlinks (text or graphics which lead you from document to document)

*interactive element, including: specially embedded programs such as: * plug-ins (downloadable sets of software that enable the user to use part of a web document.)

The base document of a home page is a file which is mostly text, containing commands ("markup") which determines how the above elements are configured. The rules governing this markup form a simple programming language called "HTML". HTML stands for "Hypertext Markup Language". As mentioned above, the Web also provides access to services. Some of these services are assembled "on the fly" by a computer program that accesses information available in some other form, and presents the information in the same format as any other web page. The Web is therefore a user interface providing access to many types of information available on the Internet.

2.5 Search Engines

A search engine is software, usually accessed on the Internet that searches a database of information according to the user's query. The engine provides a list of results that best match what the user is trying to find. Today, there are many different search engines available on the Internet, each with their own abilities and features.

The first search engine ever developed is considered Archie, which was used to search for FTP files and the first text-based search engine is considered Veronica. Currently, the most popular and well-known search engine is Google. Other popular search engines include AOL, Ask.com, Baidu, Bing, and Yahoo.

2.5.1 Search Engine Working

Web crawler, database and the search interface are the major component of a search engine that actually makes search engine to work. Search engines make use of Boolean expression AND, OR, NOT to restrict and widen the results of a search. Following are the steps that are performed by the search engine:

- The search engine looks for the keyword in the index for predefined database instead of going directly to the web to search for the keyword.
- It then uses software to search for the information in the database. This software component is known as web crawler.
- Once web crawler finds the pages, the search engine then shows the relevant web pages as a result. These retrieved web pages generally include title of page, size of text portion, first several sentences etc.
- User can click on any of the search results to open it.

2.6 Pubmed

NOTES

PubMed is the U.S. National Library of Medicine's (NLM) premiere search system for health information.

2.6.1 PubMed Content

Nearly 25 million citations including:

- Publisher supplied citations that will be analyzed to receive full indexing for MEDLINE if they are biomedical in nature
- In-process citations that have not yet been analyzed and indexed for MEDLINE
- Indexed for MEDLINE citations of articles from about 5600 regularly indexed journals; MEDLINE makes up nearly 90% of PubMed.

2.6.2 PubMed Features

- Sophisticated search capabilities, including spell checker, Advanced Search Builder, and special tools for searching for clinical topics
- Assistance in finding search terms using the MeSH (Medical Subject Heading) database of MEDLINE's controlled vocabulary
- Ability to store citation collections and to receive email updates from saved searches using PubMed's My NCBI
- Links to full-text articles, to information about library holdings, and to other NLM databases and search interfaces

2.6.3 Clinical Queries

PubMed Clinical Queries makes it easy to find articles that report applied clinical research. Click on the link from the PubMed homepage, then enter a search term in the box. Click the Search button. Click See all at the bottom of the page to return to PubMed.

Clinical Study Categories displays results by diagnosis, etiology, therapy, etc. Use the dropdown menus to change the category or scope.

Systematic Reviews displays evidence-based medicine citations including systematic reviews, meta-analyses, and guidelines.

Medical Genetics displays citations focused on diagnosis, management, genetic counseling, and related topics. Select All or a specific topic from the drop-down menu.

2.7 Public Biological databases

The past few decades computational biology has become an emerging technology and widespread applications of computers are used for analysis and modeling of biological data. Storage, retrieval and analysis of biological dataset maintained by centralized resources worldwide are the major aspect of this revolution. Dramatic increases have been observed in genomic and proteomic data as a resultant of advancement in molecular biology techniques and high throughput methods. These methods include a wide range of information such as genetic and physical genomic maps, polymorphisms, biobliographic information, molecular

NOTES

chemical properties and macromolecular sequences and structures etc., Concurrently, the fast development of biomedical knowledge, reduction in computer costs, wide spread of internet access, and the recent emergence in the field of high throughput structural and functional genomic technologies has directed to a rapid expansion of electronically available data. Such data has deposited in public archives as a various biological databases and can be used by the scientific community for querying and retrieving of necessary information. The Molecular Biology Database Collection by Micheal Galperin is a collection of several databases. The databases included in the collection that are freely available. The recent update includes 719 databases. The databases are arranged in a hierarchical classification and categorized into three different branches. National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL), and DNA Data Bank of Japan (DDBJ) are the major organizations whereas a majority of the tools and databases are being maintained.

Early in the 1960s and 1970s, Margaret Dayhoff was the first one who worked on protein sequences and was first assembled and made freely available. Now, this database named as the international Protein Information Resources (PIR). Later, SWISS-PROT sequence databases was framed and publicly announced in 1980 (Bairoch and Boeckmann, 1991). Concomitantly, various databases were initiated to flourish in

order to handle the increasing quantities of sequence data being generated worldwide.

Generally, databases can be classified into primary, secondary, and composite databases. Primary database contain the information that are directly deposited by the submitter. NCBI, EMBL, DDBJ, SWISS-PROT and PIR are the widely used primary databases for the sequences (nucleotide and proteins) and Protein Data Bank (PDB) have been universally used for protein three dimensional structures. Secondary database includes information such as conserved sequence, signature sequences, conserved secondary structure motifs and active site residues of the protein families using multiple sequence alignment. Various researchers at different individual laboratories created and maintained different databases such as SCOP, CATH, PROSITE, and eMOTIF , CDART, and protein interactions.

2.7.1 Genomic Sequence Databases – GenBank, EMBL, DDBJ

Genbank, was built and maintained by the National Center for Biotechnology Information (NCBI). It is part of the International Nucleotide Sequence Database in collaboration with the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences , acquire initially through submissions from individual laboratories

NOTES

and batch submissions from large-scale sequencing projects. NCBI's retrieval system helps to access GenBank and Entrez used to combine information from the major DNA and protein sequence databases along with mapping, taxonomy genome, protein structure and domain information, and the literature information through PubMed. The NCBI has integrated server called BLAST (Basic Local Alignment Searching Algorithm) that identifies similar

sequences for the query sequence. Up-to-now nearly 60 million sequences are available and it's growing aggressively. Genbank also provides the FTP access to the user wher they can retrieve the large set of data for their analysis from ftp://ftp.ncbi.nih.gov. Each sequence in GenBank is annotated with literature reference, size, sequence features, organism, and protein translations. The GenBank sequences are arranged under several divisions viz. EST, ENV, Plant sequences (PLN), Genome survey sequences (GSS), rodents (ROD), Sequence tagged site (STS) etc.,

GenBank entry includes a brief description of the sequence, the scientific name, and taxonomy of the source organism, bibliographic references, and a table of features. Table I.1. list major sequence databases and Table I.2. Lists different sequence divisions in Genbank.

1.7.2 Protein Sequence Databases

Swiss Protein Databank (SWISS-PROT) (Boeckmann et al., 2003), the translation of the DNA sequences in EMBL

(TrEMBL), Protein Information Resource (PIR), the Munich Information Center for Proteins (MIPS), and the 3D structures in the Protein Data Bank (PDB). Protein databases are growing faster than DNA databases.

2.8 LET US SUM UP

In this unit, you have learnt about the operating system, World Wide Web, search engines finding scientific articles. The operating system gave detail information about how our computer works and how to change our input into output. The detail application of Linux will helpful to your study like writing algorithms, developing database and command prompt methods. The World Wide Web and search engines gave the internet access and finding related articles. The database information is helpful to retrieve and store the data.

2.9 Unit-End Exercise

1. Define OS and its types?
2. How to install the software’s in Linux?
3. What are search engines?
4. Define Pub Med and its Features?
5. Give an example for primary biological databases?

NOTES

2.10 CHECK YOUR PROGRESS

1. What is operating system?
2. What is World Wide Web and note the different search engines
3. Comment on the clinical queries of pubmed and different biological databases.

2.11 ANSWERS TO CHECK YOUR PROGRESS

1**.** An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer.

2. The World-Wide Web is a collection of documents and services, distributed across the Internet and linked together by hypertext links. Search Engines: Google, AOL, Ask.com, Baidu, Bing, and Yahoo

3. Clinical Queries: PubMed Clinical Queries, Clinical Study Categories, Systematic Reviews and Medical Genetics

Biological databases: GenBank, EMBL, DDBJ, SWISS-PROT, PIR and etc..

2.12 SUGGESTED READINGS

NOTES

1. Kenneth Rosen, Douglas Host, Rachel Klee, Richard Rosinski (2007), UNIX: The Complete Reference (2nd Ed.), McGraw-Hill publication.

2. Richard Petersen (2017), Linux: The Complete Reference (6th Ed.) McGraw-Hill publication

3. Black U (1996), Computer Networks-Protocols, Standards and Interfaces, PHI.

UNIT-III

NOTES

Structure

3.1 Introduction

Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. In this unit we mainly focused on the sequence analysis, sequency

assembly, Pair wise sequence comparison and genome sequence annotation. In this unit will be helpful to analyze the sequence and their relationship. The sequence assembly and sequencing will be used to gain the knowledge about the various methods for constructing whole genome sequence. Hence, this unit elaborates the genomics and their methods.

3.2 Objectives

After going through the unit you will able to;

- Understand the sequence analysis technique
- To identify the whole genome sequencing methods
- Working procedure on pair-wise alignment
- Identify the importance and key role of genome annaotation

3.3 Genomics

Genomics is a forum for describing the development of genome-scale technologies and their application to all areas of biological investigation. That has evolved with the field that carries its name, Genomics focuses on the development and application of cutting-edge methods, addressing fundamental questions with potential interest to a wide audience.

3.3.1 The scope of Genomics

Genome Biology covers all areas of biology and biomedicine studied from a genomic and post-genomic perspective.

NOTES

It should be used in sequence analysis; bioinformatics; insights into molecular, cellular and organismal biology; functional genomics; epigenomics; population genomics; proteomics; comparative biology and evolution; systems and network biology; genome editing and engineering; genomics of disease; and clinical genomics. • Genomics including genome projects, genome sequencing, and genomic technologies and novel strategies.

• Functional genomics including transcriptional profiling, mRNA analysis, microRNA analysis, and analysis of noncoding and other RNAs using established and newly-emerging technologies (such as digital gene expression).

• Evolutionary and comparative genomics, including phylogenomics.

• Genomic technology and methodology development, with a focus on new and exciting applications with potential for significant impact in the field and emerging technologies.

• Computational biology, bioinformatics and biostatistics, including integrative methods, network biology, and the development of novel tools and techniques.

• Modern genetics on a genomic scale, including complex gene studies, population genomics, association studies, structural variation, and gene-environment interactions.

• Epigenomics, including DNA methylation, histone modification, chromatin structure, imprinting, and

chromatin remodeling.

• Genomic regulatory analysis, including DNA elements, locus control regions, insulators, enhancers, silencers, and mechanisms of gene regulation.

• Genomic approaches to understanding the mechanism of disease pathogenesis and its relationship to genetic factors, including meta-genomic and the mode and tempo of gene and genome sequence evolution.

• Medical Genomics, Personal Genomics, and other applications to human health.

• Application of Genomic techniques in model organisms that may be of interest to a wide audience.

• RNA-seq experiments without three or more biological replicates will not be accepted for differential expression-based articles.

3.4 Sequence analysis

From gene sequences to the proteins they encode to the complicated biological networks they are involved in, computational methods are available to help analyze data and formulate hypotheses. So they have focused on commonly used software packages, to attempt to encompass every detail of every program out there.

The first tools they describe are those that analyze protein and DNA sequence data. Sequence data is the most abundant type of biological data available electronically.

NOTES

While other databases may eventually rival them in size, the importance of sequence databases to biology remains central. Pairwise sequence comparison is the most essential technique in computational biology. It allows doing everything from sequence-based database searching, to building evolutionary trees and identifying characteristic features of protein families, to creating homology models. But it's also the key to larger projects, limited only by our imagination—comparing genomes, exploring the sequence determinants of protein structure, connecting expression data to genomic information, and much more.

The types of analysis that can do with sequence data are:

- Knowledge-based single sequence analysis for sequence characteristics
- Pairwise sequence comparison and sequence-based searching
- Multiple sequence alignment
- Sequence motif discovery in multiple alignments
- Phylogenetic inference

3.5 Genome Sequencing

3.5.1 History of sequencing

British biochemist named Frederick Sanger, laid the foundation for sequencing proteins. In 1955, Sanger had completed the sequence of all the amino acids in insulin. His work provided evidence that proteins consisted of

chemical entities with a specific pattern, rather than a mixture of substances.

Later, a method named as Sanger Sequencing was developed by Frederick Sanger and his colleagues in 1977, where DNA could be sequenced by generating fragments. It was the most widely used sequencing method for approximately 40 years.

NOTES

3.5.2 Whole Genome Sequencing and Sequence Assembly

A DNA sequencing reaction produces a sequence that is several hundred bases long. Gene sequences are typically thousands of bases long. The largest known gene is the one associated with Duchenne muscular dystrophy. It is approximately 2.4 million bases in length.

The basic procedure in sequencing has been to isolate genomic DNA or RNA (reverse transcribed into cDNA), and clone it into vectors (eg. plasmids, BACs) capable of stable propagation in suitable host cells such as Escherichia coli. Several cloning systems with insert sizes varying from hundreds of base pairs to megabases have been developed. The ideal clone library for genomic sequencing has the following features.

- The clones are highly redundant, covering the entire genome many times (typically 6–10).
- The clone coverage is random and not biased

NOTES

towards or against specific regions of the genome.

- The clones are stable, not subject to recombination or reorganization during the propagation process.

It should be noted that one of the major improvements of the new massively parallel sequencing technologies is that they do not rely on vector cloning prior to sequencing, and the concerns listed here are therefore not directly applicable to those technologies. After propagation, the clones are selected and the sequencing is performed. An essential feature in sequencing is the attachment of quality values to the raw sequences. The quality values indicate the likelihood of each base call being correct. In the assembly stage the quality values will help to distinguish true DNA polymorphisms from sequencing errors and match end sequences of low quality. In genomic shotgun sequencing, which typically uses a single individual DNA source, sequences sharing less than 98% identity are usually assumed to come from different regions of a genome (including different repetitive elements). In contrast, EST data is usually derived from a variety of sources representing the spectrum of polymorphisms in the original samples. These will usually include a number of erroneous polymorphism which are caused by sequencing errors

inherent in single pass sequencing, a relatively high rate of insertions and deletions, contamination by vector and linker sequences and the non-random distribution of sequence start sites in oligo(dT)-primed libraries. Therefore, the degree of identity in overlapping sequences from the same gene will often be lower in EST projects than in genomic sequencing projects. In addition, the patterns of overlapping sequences caused by alternative splice forms are different from those observed in a genomic shotgun project.

3.5.3 Shotgun sequencing

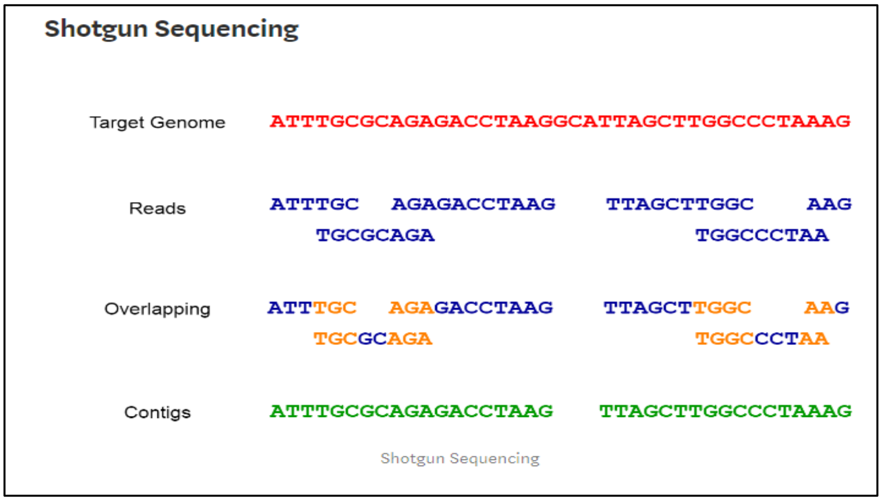
Two approaches for genome shotgun sequencing can be distinguished: whole-genome shotgun (WGS) sequencing and hierarchical shotgun sequencing.

Shotgun sequencing

Sequencing using the whole genome shotgun approach basically means that the genome is randomly broken into pieces and cloned into a sequencing vector. The inserts are subsequently processed to generate sequences of bases (referred to as reads). During the mid 1990s several groups recognized that sequence information from both ends of relatively long inserts dramatically improves the efficiency of sequence assembly. In contrast to single sequence reads from one end of the shotgun clones pairs of sequence reads from both ends have known spacing and orientation. Exact knowledge of the length of the insert is not required to utilize the

NOTES

advantages of end sequencing in assembly, but good estimates of clone length will aid the assembly immensely.



Hierarchical shotgun sequencing

The 'Hierarchical shotgun sequencing' (also referred to as 'map-based', 'BAC-based' or 'clone-by-clone') approach involves generating and organizing a set of large insert clones (typically 100–200 kb each) covering the genome (a "minimal tiling path"), followed by separate shotgun sequencing on each clone. It is possible to establish a tiling path of overlapping BAC-clones using only BAC fingerprinting technologies. The unique genome markers. (*eg.* ESTs or sequence-tagged sites (STS)) and their location in the genome map is of great help for organizing the BAC clones in the correct order. In hierarchical shotgun sequencing the sequence information is local; therefore the risk of long-range and short-range misassembly is reduced.

NOTES

Mixed strategy sequencing

A strategy that can be used on large complex genomes is the 'mixed strategy sequencing'. The technique utilizes both hierarchical and whole-genome shotgun. The method combines a light (x1) BAC clone coverage of the genome, with whole genome shotgun sequencing. The BAC clones act as a basic framework for WGS sequence assembly. The method was successfully applied to rat genome.

Reduced Representation Sequencing

A variant of WGS is "reduced representation sequencing" (RRS), where one selectively chooses subsets of the genome to avoid sequencing the (often much) larger regions that are not of interest. In, SNPs were discovered by mixing DNA from many individuals, preparing a library of appropriately sized restriction fragments, and randomly sequencing 5 clones. Here, the choice of the restriction fragments effectively selects only a small subset of the human genome. Several approaches to RRS have been employed for plant genomes: Methyl-filtration (MF) sequences uses the endogenous restriction-modification system of *E. coli* to eliminate methylated DNA inserts, the RescueMu (RM) approach focuses on the gene-rich regions which are rich in mutator transposons, and High-Cot filtration avoids repetitive and low copy sequences due to differences in the relative rates of DNA re-association. Most of the Maize and Sorghum genomes have been sequenced using MF. Many of the applications of the new high throughput sequencing

NOTES

platforms are based on various RSS strategies. This includes electrophoretic size separation to enrich for small RNA molecules; reduced representation bisulphite sequencing for genome wide methylation analysis; flow sorting of derivative translocation chromosomes for breakpointmapping; enrichment of DNA-fragments bound to specific proteins by chromatin immunoprecipitation of fixed, sheared DNA, for identification of transcription factor binding sites (CHiP-Seq); enrichment of Specific parts of the genome by multiplex PCR-amplification or by hybridization to custom made arrays , *eg.* For SNP discovery and in situ exon capture.

EST sequencing

EST (expressed sequence tag): A unique stretch of DNA within a coding region of a gene that is useful for identifying full-length genes and serves as a landmark for mapping. An EST is a sequence tagged site (STS) derived from cDNA.

An STS is a short segment of DNA which occurs but once in the genome and whose location and base sequence are known. STSs are detectable by the polymerase chain reaction (PCR), are helpful in localizing and orienting mapping and sequence data, and serve as landmarks in the physical map of the genome.

Expressed Sequence Tags (ESTs) are sequences representing genes which can originate from specific tissues. In EST-sequencing a single automated sequencing from one or both ends of cDNA-

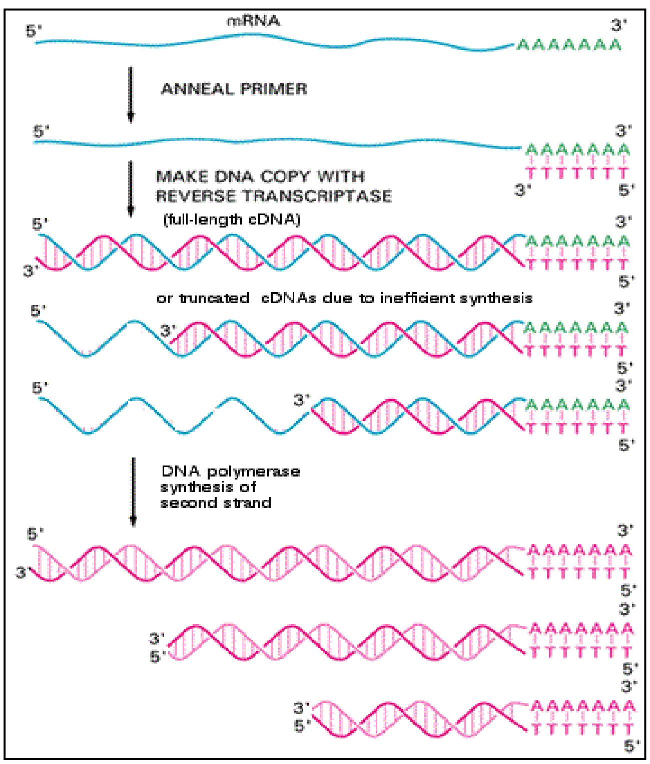
NOTES

inserts is performed. This singlepass approach is the major reason EST-sequencing is cost effective. In most cases EST sequencing projects are aimed at establishing partial sequences of transcribed genes rather than full length cDNA sequences. However, this approach features some special challenges such as common sequence motifs, alternative transcripts and paralogous genes are challenges that potentially impact the assembly quality.

Single sequence reactions from cDNA libraries of specific tissues and culture conditions

- readout of mRNA sequences present in cell

cDNAs are synthesized from mRNA mixture present in tissue of interest using reverse transcriptase.



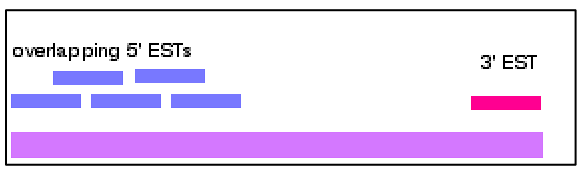
Once cDNAs are

NOTES

synthesized, they are cloned into a vector for sequencing. Usually, they are cloned in a known orientation, so that sequence can be derived from the 3' end of the message or 5' end. Since full-length cDNAs are difficult to synthesize, different cDNA copies of the same mRNA may have different sequences at their 5' ends.

- so multiple cDNAs of a particular message can give overlapping reads of mRNA sequence
- can assemble into larger "contig" of mRNA sequence by clustering (similar to genomic sequence assembly)

Clustering



Some ESTs may differ in sequence, due to variations in mRNA splicing. Sequences are assembled onto known full-length cDNAs from genbank, if known. Assembly is called a UniGene - unique gene from which ESTs are derived.

Once UniGenes determined, the EST can be used as a measure of gene expression within a tissue:

- Highly expressed genes will be represented by numerous different ESTs
- Low expression genes will only rarely show up in random library of ESTs

NOTES

But this means a few genes can be represented by most of the ESTs,

- need to sequence many more ESTs to discovered novel sequences

3.6 Sequence Assembly

Genome sequencing is a discipline that has undergone tremendous development in the past. Today 567 bacterial genomes with up to 10.5 million base pairs (Plesiocystis pacifica SIR-1) have been sequenced.

Sequence assembly is the only way to efficiently assemble sequenced fragments of DNA. However, a sufficient amount of high quality sequences are required. The assembly programs should be able to handle large data sets effectively and avoid misassemblies in the presence of large repetitive or duplicated regions and redundant sequences. To accomplish this, effective algorithms to handle large input data sets with the use of minimal computer time and memory are needed. One of the primary difficulties in computational genome assembly is to develop an algorithmic approach capable of detecting stretches of repetitive DNA without causing misassemblies. Repetitive sequences complicate assembly as different pieces of sequence can share the same repeat sequence originating from different genomic locations. Since the pieces are

NOTES

put together by searching for matching overlapping nucleotides, repeats can be put together erroneously. Typically, for shotgun data, repetitive sequences are revealed by clusters containing more overlapping reads than would be expected by chance.

In EST datasets the main difficulty is to develop an algorithmic approach that, in addition to efficient assembly, can handle highly expressed genes, paralogous genes, alternative spliceforms and chimerism in the dataset. The theoretical background for genome assembly lies in computer science, and an insight into the mathematical and theoretical background can be found in and references therein. Although pyrosequencing with a whole-genome shotgun approach has been successfully applied to bacterial genomes, the construction of high-quality assemblies with high-throughput sequencing data is still a non-trivial problem even for short genomes. At present, no approach has been proposed to directly assemble large animal or plant genomes directly from short sequences obtained using HTS. As described below the Short Read Assembly Protocol (SHRAP), however, comprises a protocol for high-throughput short read sequencing that differs in two respects from classical hierarchical sequencing approaches. This protocol however, expects read lengths much longer (200 nucleotides) than those produced by SOLiD or Solexa. The assembly methodology is based on the Euler engine introduced in 2004.

The Euler-SR assembler, specifically designed to assemble short reads, uses an updated version of the Euler engine to reduce memory requirements. The results for real Solexa reads, however, were less convincing due to the poorly understood error model and highly variable error rates across different machines and run times.

3.6.1 Basic principles of Assembly

For the majority of traditional assembly programs the basic scheme is the same, namely the overlap-layout-consensus approach.

Essentially it consists of the following steps:

• Sequence and quality data are read and the read are cleaned.

• Overlaps are detected between reads. False overlaps, duplicate reads, chimeric reads and reads with self-matches (including repetitive sequences) are also identified and left out for further treatment.

• A multiple sequence alignment of the reads is performed, and a consensus sequence is constructed for each contig layout (often along with a computed quality value for each base).

• Possible sites of misassembly are identified by combining manual inspection with quality value validation. Prior to the assembly, the electropherogram(for Sanger sequencing, images for massively parallel sequencers) for a given sequence is interpreted as a sequence of bases (a read) with associated quality values, these values reflect the log-odds score of the bases being correct. The

NOTES

basecaller PHRED is often used, however alternatives exist, *eg*. The CATS basecaller.

The reads can then be screened for any contaminant DNA such as *Escherichia coli*, cloning or sequencing vector. Low quality regions can be identified and removed. Base quality values can be used in computation of significant overlaps and in construction of the multiple alignments.

For high-throughput sequencing data, the basic proposition for SHRAP is to sample clones from the genome at high coverage, while sequencing reads from these clones at low coverage. SHRAP starts off with assembling the reads greedily to small local assemblies and subsequently to contigs on each clone. It proceeds by ordering the clones in a "clone graph", and constructing "clone contigs", which are then assembled independently. Computer simulations of the procedure show that the approach can reach a quality comparable to the current assemblies of single human chromosomes and fruit fly genomes using reads of 200nt with an error rate of not more than 1%. These are constraints that are too strict for short (Solexa or SOLiD) reads (≈ 40*bp*) and because of higher error rates challenging for real 454 reads. Furthermore, for mammalian genomes the use of a hierarchical sequencing strategy might be somewhat cumbersome. However, the use of templates might bail Solexa and SOLiD users out: In a recent study, de novo assemblies of chloroplast genomes (≈ 120 kb) were improved by aligning preassembled contigs to reference genomes. After de-

NOTES

Bruijn graph assembly of reads, small contigs were aligned to closely related chloroplast genomes. Between 67% and 98% of the contigs could be aligned to such templates. If alignment failed, sequences were scanned for similarity using BLASTN. The authors reported that successful BLASTN matches typically contained > 100 bp insertions relative to the reference genome. In the end, however, their assemblies were estimated to be 88–94% complete. The successful de novo assembly of Chloroplasts genomes with 454 reads has been shown earlier.

SOLiD, Solexa technologies allow convenient generation of mate-pair/paired-end sequences, *ie.* The ability to sequence both ends of each DNA fragment. However, in an assembly using a hybrid dataset of real 454 reads and simulated mate-pair data, about 96% of the mate-pairs did not contribute additional information and hence did not improve the assembly. Likewise, in a hybrid dataset of 454 reads and Sanger reads the vast majority of long sequences did not improve the assembly substantially, measured by N50 contig size. Hence, the authors concluded that hybrid protocols should be reviewed critically. Despite those simulation results, the latter method has already been shown to work quite well in practice, and one area wheremate-pair/paired-end sequencing should improve the analysis dramatically is for the detection of breakpoints related to structural rearrangements, *eg.* Deletions, duplications, inversions and translocations

3.7 Pairwise Sequence Comparison

NOTES

Pair-wise alignment produces an optimal alignment using a scoring matrix for the given two sequences

Description

Pairwise alignment is an arrangement in columns of 2 sequences, highlighting their similarity.

Input

Protein sequence is given as input.

G A A T T C

G A T T A

Procedure

1. Go to http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html
2. Paste the protein sequence and click compute parameters.
3. The result page is retrieved.

Output

NOTES

```
########################################
# Program: needle
# Rundate: Fri 13 May 2016 12:03:53
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20160513-120351-0311-34620455-oy.asequence
#    -bsequence emboss_needle-I20160513-120351-0311-34620455-oy.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 6
# Identity:       3/6 (50.0%)
# Similarity:     3/6 (50.0%)
# Gaps:           1/6 (16.7%)
# Score: 7.0
#
#
#=======================================

EMBOSS_001         1 GAATTC      6
                      .|||.
EMBOSS_001         1 -GATTA      5
```

<u>Output for Pair wise Alignment</u>

3.8 Annotating and analyzing genome sequences

DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary

Genome annotation consists of three main steps:

1. Identifying portions of the genome that do not code for proteins
2. Identifying elements on the genome, a process called gene prediction
3. Attaching biological information to these elements

NOTES

Automatic annotation tools attempt to perform these steps via computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

A simple method of gene annotation relies on homology based search tools, like BLAST, to search for homologous genes in specific databases; the resulting information is then used to annotate genes and genomes. However, as information is added to the annotation platform, manual annotators become capable of deconvoluting discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g. Ensembl) rely on curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- vgene structure
- coding regions
- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. Proteogenomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.

A variety of software tools have been developed to permit scientists to view and share genome annotations; for example, MAKER.

Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals (The 100,000 Genomes Project,UK)and several model organisms are largely complete. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".

Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic

NOTES

means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

1. Encyclopedia of DNA elements (ENCODE)
2. Entrez Gene
3. Ensemble
4. GENCODE
5. Gene Ontology Consortium
6. GeneRIF
7. RefSeq
8. Uniprot
9. Vertebrate and Genome Annotation Project (Vega)

3.9 LET US SUM UP

In this unit, you have learnt about the genome sequencing assembly, analayis and annotation. This knowledge would make understand how the sequence is constructed and various analyzing techniques. The sequence assembly and annotation would pave the key attention to the student for the structural and functional aspect of student. The sequencing techniques are learnt.

3.10 Unit-End Exercise

1. Define genomics and its scope?
2. What are the various types of analysis that can do with sequence data?
3. How Whole Genome Sequencing is done?

3.11 CHECK YOUR PROGRESS

1. What is Sequence analysis?
2. Comment on different types of sequencing methods
3. Note down the two types of annotation

3.12 ANSWERS TO CHECK YOUR PROGRESS

NOTES

1**.** Sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

2. Shotgun sequencing, Hierarchical shotgun sequencing, mixed strategy sequencing, Reduced Representation Sequencing and EST sequencing

3. Structural and functional Annotation

3.13 SUGGESTED READINGS

1. David Mount, (2004), “Bioinformatics: Sequence and Genome Analysis”; Cold Spring harbor laboratory Press, US Revised Edition.

2. Baxevanis, A.D. and Francis Ouellellette, B.F. (2011) “Bioinformatics –a practical guide to the analysis of Genes and Proteins”; John Wiley & Sons, UK, Third Edition.

3. Bergeron B. (2003). Bioinformatics Computing - The Complete Practical Guide to Bioinformatics for Life Scientists, by Prentics- Hall, Inc., New Jersey 07458, USA, 1st edition, ISBN :81-203-2258-4.

NOTES

BLOCK 2: Database

UNIT-IV

Structure

4.1 Introduction

Sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. In this unit we will discuss about the sequence related databases, BLAST and FASTA analysis, and multifunctional tool analysis. This unit is mainly helpful to predict the function of unkown sequence. The

genbank database information is used to gain the knowledge about various tools and sequence database annotation and etc..

4.2 Objective

- Sequence submission and retrieval form Genbank will be learnt.
- To identify the methods and procedure of BLAST and FASTA tools
- Know about the multifunctional tools for sequence analysis

4.3 Genbank sequence queries against biological databases

Genbank, was built and maintained by the National Center for Biotechnology Information (NCBI). It is part of the International Nucleotide Sequence Database in collaboration with the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences, acquire initially through submissions from individual laboratories and batch submissions from large-scale sequencing projects. NCBI’s retrieval system helps to access GenBank and Entrez used to combine information from the major DNA and protein sequence databases along with mapping, taxonomy genome, protein structure and domain information, and the literature information through PubMed (McEntyre and Lipman, 2001). The NCBI has integrated server called BLAST (Basic Local Alignment Searching Algorithm) that identifies similar sequences for the query

NOTES

sequence. Up-to-now nearly 60 million sequences are available and it's growing aggressively. Genbank also provides the FTP access to the user wher they can retrieve the large set of data for their analysis from ftp://ftp.ncbi.nih.gov. Each sequence in GenBank is annotated with literature reference, size, sequence features, organism, and protein translations. The GenBank sequences are arranged under several divisions *viz.* EST, ENV, Plant sequences (PLN), Genome survey sequences (GSS), rodents (ROD), Sequence tagged site (STS) *etc*.,

GenBank entry includes a brief description of the sequence, the scientific name, and taxonomy of the source organism, bibliographic references, and a table of features (http://www.ncbi.nlm.nih.gov/collab/FT/index.html). Table 1. lists different sequence divisions in Genbank.

Table 1. Different division under which sequences are made available in GenBank

Code	Description
PRI	primate sequences
ROD	rodent sequences
MAM	other mammalian sequences
VRT	other vertebrate sequences
INV	invertebrate sequences
PLN	plant, fungal, and algal sequences
BCT	bacterial sequences
VRL	viral sequences
PHG	bacteriophage sequences

NOTES

SYN	synthetic sequences
UNA	unannotated sequences
EST	EST sequences (expressed sequence tags)
PAT	patent sequences
STS	STS sequences (sequence tagged sites)
GSS	GSS sequences (genome survey sequences)
HTG	HTGS sequences (high throughput genomic sequences)
HTC	HTC sequences (high throughput cDNA sequences)
ENV	Environmental sampling sequences

4.4 Basic Local Alignment Search Tool (BLAST)

NIH has developed the BLAST tool. It uses statistical methods to estimate hits for their significance. This program recognizes sequences in the database that share common words of a pre-set size (k-tuple) and these matching words are expanded to identify un-gapped common segments between two sequences. A brief note on the alignment statistics and implementation of the BLAST search method is available at National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

There are many variants such as blastx or tblastn can efficiently used to search DNA sequence or translated sequences. BLAST search collects closely related sequences with high sequence similarity. Divergent sequences and an initial search produces no or very few hits can be efficiently managed by Higher order PAM (PAM250) or low order BLOSUM (BLOSUM40) matrices. The

NOTES

weaker similarities can be matched using PSIBLAST by an iterative position specific scoring matrix. In the first iteration PSIBLAST also function as BLAST. However, once sequences with specified E-value cutoff are identified a position specific matrix (PSSM) can be calculated in order to identify other related sequences. A more elegant note on the implementation algorithm is described at the website http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html.

4.5 FASTA

The shared regions in two sequences and score sequences in a database for homology can rapidly be identified by FASTA program (Pearson, 1998). The final output consists of a rank ordered list of sequences and alignment between sequences. The high similarity regions among the sequences are identified by segments with high frequency of conserved letters (ktup). A general value set for protein sequence search is 2 and it will look for pairs of conserved alphabets. Needleman-Wunch-Sellers algorithm is used to calculate the optimized score. The FASTA can also be used for DNA and protein sequence search. A web interface to FASTA tool that can search sequence database is available at http://fasta.bioch.virginia.edu/. Different databases such as PDB (http://www.rcsb.org) and GENOME DB (http://fasta.genome.jp) also provide FASTA search facility.

4.6 Multifunctional Tools for Sequence Analysis

4.6.1 NCBI SEALS

The NCBI SEALS project aims to develop a Perl-based command-line environment for Systematic Analysis of Lots of Sequences. SEALS are far from a fully automated genome analysis tool, and it isn't intended to be. What SEALS does provide is an enhancement to the command-line environment on UNIX systems. It is composed of a large suite of scripts with a variety of useful functions: converting file formats, manipulating BLAST results and FASTA files, database retrieval, piping files into Netscape, and a host of other features that make your data easier to look at without requiring a resource-sucking GUI. UNIX system is probably most useful for those who are already UNIX aficionados. Before you write a script to process a lot of sequences, check to see if the process you want has been implemented in SEALS.

4.6.2 The Biology Workbench

The San Diego Supercomputing Center offers access to sequence-analysis tools through the Biology Workbench. This resource has been freely available to academic users in one form or another since 1995. Users obtain a login and password at the SDSC site, and work sessions and data can be saved securely on the server.

The Biology Workbench offers keyword and sequence-based searching of nearly 40 major sequence databases and over 25 whole genomes. Both BLAST and FASTA are implemented as search and alignment tools in the Workbench, along with several local and global alignment tools, tools for DNA sequence translation, protein sequence feature analysis, multiple sequence alignment, and phylogenetic tree drawing. The Workbench group

NOTES

has not yet implemented profile tools, such as MEME, HMMer, or sequence logo tools, although PSI-BLAST is available for sequence searches.

Although its interface can be somewhat cumbersome, involving a lot of window scrolling and button clicking, the Biology Workbench is still the most comprehensive, convenient, and accessible of the web-based toolkits. One of its main benefits is that many sequence file formats are accepted and translated by the software. Users of the Workbench need never worry about file type incompatibility and can move seamlessly from keyword-based database search, to sequence-based search, to multiple alignments, to phylogenetic analysis.

4.6.3 DoubleTwist

Another entry into the sequence analysis portal arena is DoubleTwist at http://doubletwist.com. This site allows you to submit a sequence for comparison to multiple databases using BLAST. It also provides "agents" that monitor databases for new additions that match a submitted sequence and automatically notifies the user. These services, as well as access to the EcoCyc pathways database and to an online biology research protocols database, are free with registration at the site at the time of this writing.

4.7 Let us sum up

GenBank is a comprehensive database that restrains more than 1, 65,000 named organisms DNA sequences which is useful to student for gaining knowledge to every organisms. The

relationship between the two sequences is identified through BLAST and FASTA analysis. Student can know about the function of unknown sequence they learn to analyze the sequence.

4.8 Unit-End Exercise

1. Define Basic Local Alignment Search Tool (BLAST).
2. Define FASTA.
3. Types Multifunctional Tools for Sequence Analysis?

4.9 Check your progress

1. What is Genbank?
2. Define BLAST and FASTA
3. List out the multifunctional tools for sequence analysis

4.10 Answers to check your progress

1. GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences, acquire initially through submissions from individual laboratories and batch submissions from large-scale sequencing projects.
2. BLAST search collects closely related sequences with high sequence similarity and the shared regions in two sequences and score sequences in a database for homology can rapidly be identified by FASTA program.
3. NCBI SEALS, The Biology Workbench and DoubleTwist

4.11 SUGGESTED READINGS

1. Scott Markel (2003) “Sequence Analysis in a Nutshell – A Guide to Common Tools & Databases”; O'Reilly; 1 edi tion, ISBN-13: 978-0596004941

2. Baxevanis, A.D. and Francis Ouellellette, B.F. (2011) “Bioinformatics –a practical guide to the analysis of Genes and Proteins”; John Wiley & Sons, UK, Third Edition.

UNIT-V

NOTES

Structure

5.1 Introduction

Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column. In this unit, we are mainly focusing the multiple sequence alignment and phylogenetic alignment. Hence, it is essential to find the ancestor of unknown

sequence. Based on this identification we can easily predict the characteristic and function of a sequence.

5.2 Objectives

- To perform multiple sequence alignment with various tools
- To construct phylogenetic tree between the sequences
- Identify the motif and profile.

5.3 Multiple-Sequence Alignment

Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column. The principle of the BLAST or FASTA can identify closely related sequences but it is not provide information about conserved blocks of residues in a protein family. The main application of creating multiple sequence alignment of several protein sequences is to identify the significant conserved regions and in many cases to understand function. Heuristic methods based on progressive-alignment (ClustalW, TCOFFEE), simultaneous alignment of all sequences (MSA), using iterative strategies (Prrp) are the major contributors to produce multiple global sequence alignments (Higgins and Sharp, 1988).

5.3.1 ClustalW

ClustalW is a multiple sequence alignment tool for protein or DNA sequences. The closely related sequences are added and more evolutionarily diverged sequences are added in this

NOTES

algorithm. Two different steps to produce a complete multiple alignment are of (1) Identification of closely related sequences among the input, (2) Ordering these sequences based on pair-wise similarity score, (3) Alignment construction using those sequence pairs with highest similarity score, (4) Addition of sequences in a tree-order manner to create alignments. EMBL website contains a web interface at http://www.ebi.ac.uk/clustalw/.

5.3.2 TCoffee

TCoffee is a multiple alignment tool based on progressive-alignment strategy on a scoring matrix. A global and local pair-wise alignment of all sequences is created as a first step and weights for individual residue pairs are assigned based on the percentage identity of aligned residues. Further, using this library of primary alignments, an extended library of alignments has generated by considering all possible primary alignments. A correct alignment is generated using dynamic programming. A correct multiple alignment should characterize a structural alignment among input sequences. BaliBase database has been used to test the TCoffee accuracy of alignment (Thomopson *et al.*, 1999). T-Coffee program can be accessed at http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi. Both TCoffee and Clustal methods are included in the BioPerl package (http://www.bioperl.org).

5.4 Phylogenetic alignment

Phylogenetics is the study of evolutionary relationships. The evolutionary history

Deduced from phylogenetic analysis is usually represented as branching, tree like diagrams that represent an estimated pedigree of the inherited relationships among molecules ("gene trees"), organisms. Phylogenetics is sometimes called as cladistics. Sequences are outshined morphological and other organismal characters as the most popular form of data for phylogenetic or cladistic analysis. Numerous phylogenetic algorithms, procedures, and computer programs have been invented and their reliability and practicality are mainly on the structure and size of the data.

5.4.1 Steps involved in Phylogenetic analysis

1. Alignment
2. Determining the substitution model
3. Tree building
4. Tree evaluation

Alignment (Building the data model)

Multiple sequence alignments is the major part of phylogenetic sequence alignments. The alignment based positions are referred as "sites". These sites are equivalent to "characters".

Determining the substitution model

The substitution model is also very important as like as alignment and tree building. This model manipulates the alignment and tree building. The model of substitution can be generated either from particular bases or from the relative rate of overall substitution among different sites in sequences.

NOTES

Tree building methods

Tree building methods can be arranged into character based Vs distance based methods. Character based method can be derived as the optimization of the distribution of the actual data patterns for each character. The distance based methods compute the pairwise distances according to some measurement to derive trees.

Tree evaluation

Several procedures are available that evaluate the phylogenetic signal in the data and the robustness of trees. There are tests of data signal versus randomized data (skewness and permutation tests) is the most popular method to evaluate the constructed phylogentic tree. The likelihood ratio test provides a means of evaluating both the substitution model and the tree.

5.5 Motifs

A related set of nucleotides or residues may possess conserved patterns may have specific functions. These conserved patterns are known as motifs. The transcription factors may bind to specific nucleotide motifs in DNA sequences. The initial step of this process is to collect a set of sequences with common function. These sequences may belong to the same protein family or may be diverse with the common function. PROSITE is a widely used database of patterns or motifs with details of function and sequence contain these patterns. Currently it contains 1309 patterns. Commonly occurring patterns are expressed with IUPAC single letter amino acid code. Square brackets represent positions where

more than one amino acid. Excluding a set of amino acids at a particular position represented in a curly braces. Minimum and maximum repetition of residues is represented in numbers inside a bracket. PROSITE patterns can be downloaded in http://www.expasy.ch/prosite/ from ExPasy website. Patterns can be identified for a given sequence in PROSITE database using search engine “ScanProsite”. Two different methods viz. physical-chemical properties and scoring the occurrence of amino acids with higher rate of substitution were used to score the conservation of residues in a particular position.

NOTES

5.6 Profile

Profile is a quantitative or qualitative method of describing a motif. A profile can be expressed in its most rudimentary form as a list of the amino acids occurring at each position in the motif. Early profile methods used simple profiles of this sort; however, modern profile methods usually weight amino acids according to their probability of being observed at each position

5.7 Let us sum up

In this unit, you have learnt about the multiple sequence and phylogenic alignment methods and procedures. The concept such phylogenic alignment and tree construction would have made you to generating Phylogenetic tree. This content might play very important role in Phylogenetic tree construction.

5.8 Unit-End Exercise

1. Define Multiple-Sequence Alignment
2. Explain Tree building methods and their types?

NOTES

3. What are the Steps involved in Phylogenetic analysis

5.8 Check your progress

1. Define Multiple Sequence alignment
2. Explain Phylogenetic analysis

5.9 Answers to check your progress

1. Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column.
2. Phylogenetics is the study of evolutionary relationships. The evolutionary history deduced from phylogenetic analysis is usually represented as branching, tree like diagrams that represent an estimated pedigree of the inherited relationships among molecules ("gene trees"), organisms.

5.10 SUGGESTED READINGS

1. David Mount, (2004), "Bioinformatics: Sequence and Genome Analysis"; Cold Spring harbor laboratory Press, US Revised Edition.
2. Page, R. D. M. and Holmes, E.C. (1998) "Molecular Evolution A Phylogenetic Approach"; Blackwell Scientific.
3. Philippe Lemey, Marco Salemi and Anne-Mieke Vandamme (2009), "The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing"; 2nd Edition, Cambridge University Press

UNIT-VI

Structure

NOTES

6.1 Introduction

Proteomics is the analysis of the entire protein complement of a cell, tissue, or organism under a specific, defined set of conditions. In this unit, we cover the protein structure determination and prediction. The determination of protein structure denotes how the protein structures were solved by experimentally. The PDB and SWISSPROT database have the information about protein structure and function. Homology modeling is used predict the protein structure by bioinformatics approaches. This unit gives a vast attention about the protein structures to the student. This unit might be helpful to the learner to understand the protein structure and function.

6.2 Objectives

1. Retrieve the protein structure from PDB and SWISSPROT database
2. To understand the structure determination of protein
3. To predict the secondary and 3D structure of protein
4. Analyze the biochemical pathway of rel;ated protein

6.3 Protein Data Bank

The Protein Data Bank is the single worldwide repository of large molecules of proteins and nucleic acids. It is available at

http://www.pdb.org/. The PDB contains 1,25,180 protein structures and 2,895 nucleic acid structures. This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that help students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

6.4 SWISS-PROT

SWISS-PROT is a protein knowledge based database that contains amino acid sequence, protein name, taxonomic data, and citation information. It also offers cross-references to external data collections such as the fundamental DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence database, 2D and 3D protein structure databases, post translational modification (PTM) databases, various protein domain and family characterization databases, species-specific data collections, disease databases and variant databases. SWISS PROT is steadily being improved by the addition of a number of features from OMIM GeneLynx, GeneCards, Genew as well as to many gene-specific mutation databases. SWISS-PROT and TrEMBL can be obtained by anonymous FTP from the ExPASy server ftp://expasy.org and EBI server ftp://ebi.ac.uk/pub/. Further, user can obtain weekly updates and complete data sets in various formats from http://www.expasy.org/sprot/download.html.

NOTES

6.5 Biochemical Pathway Databases

6.5.1 WIT and KEGG

WIT is a metabolic pathway reconstruction resource; that is, the curators of WIT are attempting to reconstruct complete metabolic pathway models for organisms whose genomes have been completely sequenced. WIT currently contains metabolic models for 39 organisms. The WIT models include far more than just metabolism and bioenergetics; they range from transcription and translation pathways to transmembrane transport to signal transduction.

The General Search function allows you to search all of WIT, or subsets of organisms. This type of search is based on keywords, using Unix-style regular expressions to find matches. There is also a similarity search function that allows you to search all the open reading frames (ORFs) of a selected organism for sequence pattern matches, using either BLAST or FASTA. Pathway queries require you to specify the names of metabolites and/or specific EC enzyme codes. Enzyme queries allow you to specify an enzyme name or EC code, along with location information such as tissue specificity, cellular compartment specificity, or organelle specificity. In all except the regular-expression searches, the keywords are drawn from standardized metabolic vocabularies. WIT searches require some prior knowledge of these vocabularies when you submit the query. WIT was primarily designed as a tool to aid its developers

NOTES

in producing metabolic reconstructions, and documentation of the vocabularies used may not always be sufficient for the novice user. WIT is relatively text-heavy, although at the highest level of detail, metabolic pathway diagrams can be displayed.

Another web-based metabolic reconstruction resource is KEGG, which provides its metabolic overviews as map illustrations, rather than text-only, and can be easier to use for the visually-oriented user. KEGG also provides listings of EC numbers and their corresponding enzymes broken down by level, and many helpful links to sites describing enzyme and ligand nomenclature in detail. The LIGAND database, associated with KEGG, is a useful resource for identifying small molecules involved in biochemical pathways. Like WIT, KEGG is searchable by sequence homology, keyword, and chemical entity; you can also input the LIGAND ID codes of two small molecules and find all of the possible metabolic pathways connecting them.

6.5.2 PathDB

PathDB is another type of metabolic pathway database. While it contains roughly the same information as KEGG and WIT—identities of compounds and metabolic proteins, and information about the steps that connect these entities—it handles information in a far more flexible way than the other metabolic databases. Instead of limiting searches to arbitrary metabolic pathways and describing pathways with preconceived images, PathDB allows you to find any set of connected reactions that link point A to point B, or compound A to compound B.

NOTES

PathDB contains, in addition to the usual search tools, a pathway visualization interface that allows you to inspect any selected pathway and display different representations of the pathway. The PathDB developers plan to incorporate metabolic simulation into the user interface as well, although those features aren't available at the time of this writing.

6.6 Predicting Protein Structure and function from sequence

The amino acid sequence of a protein contains interesting information in and of itself. A protein sequence can be compared and contrasted with the sequences of other proteins to establish its relationship, if any, to known protein families, and to provide information about the evolution of biochemical function. However, for the purpose of understanding protein function, the 3D structure of a protein is far more useful than its sequence.

The key property of proteins that allows them to perform a variety of biochemical functions is the sequence of amino acids in the protein chain, which somehow uniquely determines the 3D structure of the protein. Given 20 amino acid possibilities, there are a vast number of ways they can be combined to make up even a short protein sequence, which means that given time, organisms can evolve proteins that carry out practically any imaginable purpose.

Each time a particular protein chain is synthesized in the cell, it folds together so that each of the critical chemical groups for that particular protein's function is brought into a precise geometric arrangement. The fold assumed by a protein sequence doesn't vary.

Every occurrence of that particular protein folds into exactly the same structure.

Despite this consistency on the part of proteins, no one has figured out how to accurately predict the 3D structure that a protein will fold into based on its sequence alone. Patterns are clearly present in the amino acid sequences of proteins, but those patterns are degenerate; that is, more than one sequence can specify a particular type of fold. While there are thousands upon thousands of ways amino acids can combine to form a sequence of a particular length, the number of unique ways that a protein structure can organize itself seems to be much smaller. Only a few hundred unique protein folds have been observed in the Protein Data Bank. Proteins with almost completely nonhomologous sequences nonetheless fold into structures that are similar. And so, prediction of structure from sequence is a difficult problem.

6.7 Determination of structure

The protein structure was determined by X-ray Crystallography and NMR Spectroscopy methods

6.7.1 X-ray Crystallography

Linus Pauling and Robert Corey began to use x-ray crystallography to study the atomic structures of amino acids and peptides. In experiments that began in the early 1950s, John Kendrew determined the structure of a protein called myoglobin, and Max Perutz determined the structure of a similar protein called hemoglobin. Both proteins are oxygen transporters, easily isolated in large quantities from blood and readily crystallized. Obtaining

NOTES

x-ray data of reasonably high quality and analyzing it without the aid of modern computers took several years. The structures of hemoglobin and myoglobin were found to be composed of high-density rods of the dimensions expected for Pauling's proposed alpha helix. Two years later, a much higher-quality crystallographic data set allowed the positions of 1200 of myoglobin's 1260 atoms to be determined exactly. The experiments of Kendrew and Perutz paved the way for x-ray crystallographic analysis of other proteins. In x-ray crystallography, a crystal of a substance is placed in an x-ray beam. X-rays are reflected by the electron clouds surrounding the atoms in the crystal. In a protein crystal, individual protein molecules are arranged in a regular lattice, so x-rays are reflected by the crystal in regular patterns. The x-ray reflections scattered from a protein crystal can be analyzed to produce an electron density map of the protein (see Figure 10-1: images are courtesy of the Holden Group, University of Wisconsin, Madison, and Bruker Analytical X Ray Systems). Protein atomic coordinates are produced by modeling the best possible way for the atoms making up the known sequence of the protein to fit into this electron density. The fitting process isn't unambiguous; there are many incrementally different ways in which an atomic structure can be fit into an electron density map, and not all of them are chemically correct. A protein structure isn't an exact representation of the positions of atoms in the crystal; it's simply the model that best fits

both the electron density map of the protein and the stereo chemical constraints that govern protein structures.

In order to determine a protein structure by x-ray crystallography, an extremely pure protein sample is needed. The protein sample has to form crystals that are relatively large (about 0.5 mm) and singular, without flaws. Producing large samples of pure protein has become easier with recombinant DNA technology. However, crystallizing proteins is still somewhat of an art form. There is no generic recipe for crystallization conditions (e.g., the salt content and pH of the protein solution) that cause proteins to crystallize rapidly, and even when the right solution conditions are found, it may take months for crystals of suitable size to form.

Many protein structures aren't amenable to crystallization at all. For instance, proteins that do their work in the cell membrane usually aren't soluble in water and tend to aggregate in solution, so it's difficult to solve the structures of membrane proteins by x-ray crystallography. Integral membrane proteins account for about 30% of the protein complement (proteome) of living things, and yet less than a dozen proteins of this type have been crystallized in a pure enough form for their structures to be solved at atomic resolution.

6.7.2 NMR Spectroscopy

An increasing number of protein structures are being solved by nuclear magnetic resonance (NMR) spectroscopy. NMR detects atomic nuclei with nonzero spin; the signals produced by these nuclei are shifted in the magnetic field depending on their

NOTES

electronic environment. By interpreting the chemical shifts observed in the NMR spectrum of a molecule, distances between particular atoms in the molecule can be estimated. To be studied using NMR, a protein needs to be small enough to rotate rapidly in solution (on the order of 30 kilodaltons in molecular weight), soluble at high concentrations, and stable at room temperature for time periods as long as several days.

Analysis of the chemical shift data from an NMR experiment produces a set of distance constraints between labeled atoms in a protein. Solving an NMR structure means producing a model or set of models that manage to satisfy all the known distance constraints determined by the NMR experiment, as well as the general stereochemical constraints that govern protein structures. NMR models are often released in groups of 20 -40 models, because the solution to a structure determined by NMR is more ambiguous than the solution to a structure determined by crystallography. An NMR average structure is created by averaging such a group of Models. Depending on how this averaging is done, stereochemical errors may be introduced into the resulting structure, so it's generally wise to check the quality of average structures before using them in modeling.

6.8 Feature Detection

Features in protein sequences represent the details of the protein's function. These usually include sites of post-translational modifications and localization signals. Post- translational modifications are chemical changes made to the protein after it's

transcribed from a messenger RNA. They include truncations of the protein (cleavages) and the addition of a chemical group to regulate the behavior of the protein (phosphorylation, glycosylation, and acetylation are common examples).

Localization or targeting signals are used by the cell to ensure that proteins are in the right place at the right time. They include nuclear localization signals, ER targeting peptides, and transmembrane helices. Protein sequence feature detection is often the first problem in computational biology tackled by molecular biologists. Unfortunately, the software tools used for feature detection aren't centrally located. They are scattered across the Web on personal or laboratory sites, and to find them you'll need to do some digging using your favorite search engine.

One collection of web-based resources for protein sequence feature detection is maintained at the Technical University of Denmark's Center for Biological Sequence Analysis Prediction Servers (http://www.cbs.dtu.dk/services/). This site provides access to server-based programs for finding (among other things) cleavage sites, glycosylation sites, and subcellular localization signals. These tools all work similarly; they search for simple sequence patterns, or profiles, that are known to be associated with various post-translational modifications. The programs have standardized interfaces and are straightforward to use: each has a submission form into which you paste your sequence of interest, and each returns an HTML page containing the results.

6.9 Secondary Structure Prediction

NOTES

Secondary structure prediction is often regarded as a first step in predicting the structure of a protein. Secondary structure prediction methods can be divided into alignment-based and single sequence-based methods. Alignment-based secondary structure prediction is quite intuitive and related to the concept of sequence profiles. In an alignment-based secondary structure prediction, the investigator finds a family of sequences that are similar to the unknown. The homologous regions in the family of sequences are then assumed to share the same secondary structure and the prediction is made not based on one sequence but on the consensus of all the sequences in the set. Single sequence-based approaches, on the other hand, predict local structure for only one unknown.

Modern methods for secondary structure prediction exploit information from multiple sequence alignments, or combinations of predictions from multiple methods, or both. These methods claim accuracy in the 70 -77% range. Many of these programs are available for use over the Web. They take a sequence (usually in FASTA format) as input, execute some procedure on them, then return a prediction, usually by email, since the prediction algorithms are compute-intensive and tend to be run in a queue. The following is a list of most commonly used methods:

6.9.1 PHD

PHD combines results from a number of neural networks, each of which predicts the secondary structure of a residue based on the local sequence context and on global sequence characteristics (protein length, amino acid frequencies, etc.) The final prediction,

then, is an arithmetic average of the output of each of these neural networks. Such combination schemes are known as jury decision or (more colloquially) winner-take-all methods. PHD is regarded as a standard for secondary structure prediction.

6.9.2 PSIPRED

PSIPRED combines neural network predictions with a multiple sequence alignment derived from a PSI-BLAST database search. PSIPRED was one of the top performers on secondary structure prediction at CASP 3.

6.10 Predicting 3D strcture and protein modeling

The main goal of homology modeling is to predict a structure from its sequence based on the similarity and identity. The target sequence will be retrieved from the NCBI. Further, the target sequences will be compared with known structures stored in the PDB (using BLAST). In case if any protein structures with 50% identical residues with the target sequence, then the sequences will be aligned and finally we arrive at model structure. In practice, homology modeling is a multistep process that can be divided into following steps:

1. Template identification and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side chain modeling
6. Model optimization

NOTES

7. Model validation

6.10.1 Template identification and initial alignment

The percentage of identity between the target sequence and a possible template should be high enough with simple sequence alignment programs such as BLAST (Altschul *et al.*, 1990) or FASTA (Pearson, 1990). Two different matrices namely residue exchange matrix and alignment matrix can be used for this function.

6.10.2 Alignment Correction

In some cases, it is very difficult to align two sequences in a region where the percentage of sequence identity is very low. In such cases many programs (e.g. CLUSTALW) can be used to align the sequences. It provides some additional information on sequence alignment. Multiple sequence alignments are outstanding technique in homology modeling, for example, to place deletions (missing residues in the model) or insertions (additional residues in the model) in sequences are highly divergent region.

6.10.3 Backbone Generation

After aligning the sequence, we can initiate the actual model building. Two different procedures are available: 1) The backbone coordinates (N, C, Cα and O) only be copied if two aligned residue differ, 2) the two aligned residues are same then side chains also can be included.

6.10.4 Loop Modeling

The target and template alignment structure contains gaps in majority of cases. These gaps may come from either in the

model sequence or in the template sequence. In such case, there two different methods such as knowledge based and energy based methods are used for the loop modeling.

Knowledge based: It searches the PDB for known loops with endpoints that match the residue between with the loop has to be inserted, and simply copies the loop information (Eg: 3D-Jigsaw, Insight, Modeller, Swiss-model).

Energy based: In this case an energy function has used to judge the quality of loop by *ab inito* fold prediction. Further, the energy was minimized using Monte Carlo or molecular dynamics simulation techniques to get the best loop conformation.

6.10.5 Side-Chain Modeling

The side-chain conformations (rotamers) of residues are conserved in structurally similar proteins. They often have similar χ_1 – angles (i.e., the torsion angle about the C_α - C_β bond). Thus, it is possible to copy the conserved residues entirely from the template to the model. It produces higher accuracy structure than by copying just the backbone and repredicting the side chains.

6.10.6 Model Optimization

The backbone correction is the critical step in model building since their turn depends on the rotamers and their packing. Model optimization is a significant procedure since there are more possibilities to mislead the co-ordinates of the generated structure. Two different ways to achieve the precision:

1. Qunatum force fields: Protein force fields should be fast to handle large molecules efficiently. Thus energies are normally

NOTES

expressed as function of the atomic nuclei positions. More accurate electrostatics has applied to derive the classical force fields.

2. Self-paramaterizing force fields: The large extent of force field parameters affects the force field accuracy. These force fields are usually attained from quantum chemical calculations (e.g., Van der Waals radii, atomic charges).

However, the most straightforward approach to model optimization is to run a MD simulation of the model. This process mimics true folding process. Thus, the model will complete its folding and “home in” to the true structure during the simulation.

Model validation

In most of the cases the generated models may contain errors. The number of errors mainly depends on two values:

1. The percentage sequence identity between target and template. If the similarity is > 90% than the accuracy of the model can be compared to crystallographically determined structures. In case of 50% - 90% identity, the rms error in the modeled co-ordinated can be > 1.5 Å. If the sequence identity is < 25%, the alignment turns out to be the main bottleneck for homology modeling.

There are two different ways to estimate errors in a structure:

1. Calculating the models energy based on a force field: This methods check the bond length and bond angle are within the normal ranges.

2. Determination of normality of the modeled structure on how well a given characteristic resembles the same characteristic in real

structures. Most of them are mainly depend on the interatomic distances and contacts.

6.11 Let Us Sum Up

In this unit you have learnt about the structure prediction, determination, and pathway analysis, protein information and modeling. This knowledge might a better understanding about the protein structure. The databases and tools would be helpful to identify the particular protein and their sequence, structure and functional annotation. This information is also used to predict the structure of a protein.

6.12 Unit-End Exercise

1. Explain the Steps involved in homology modeling?
2. Define Protein Data Bank.
3. Explain the various types of Biochemical Pathway Databases?

6.13 Check your progress

1. Define PDB and Swiss prot
2. List out the Biochemical Pathway Databases
3. Explain the structure determination methods
4. List out the various steps of protein modelling

6.14 Answers to check your progress

1. Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that help students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to

health and disease. SWISS-PROT is a protein knowledge based database that contains amino acid sequence, protein name, taxonomic data, and citation information.

2. WIT, PathDB and KEGG
3. X-ray Crystallography and NMR Spectroscopy
4. Template identification and initial alignment, Alignment correction, Backbone generation, Loop modeling, Side chain modeling, Model optimization, Model validation

6.15 SUGGESTED READINGS

1. Bourne, P. E. & Weissig, H. (2003), "Structural bioinformatics"; Wiley-Liss

2. George H. Stout, Lyle H. Jensen (1989), "X-Ray Structure Determination": John Wiley & Sons

3. Leach, AR (2001), "Molecular Modeling – Principles and Applications"; Second Edition, Prentice Hall, USA

4. Teresa K. Attwood, David J. Parry-Smith (1999), Introduction to bioinformatics, Pearson Education

BLOCK 3: Introduction to Biostatistics

UNIT-VII

NOTES

Structure

NOTES

7.9 Suggested readings

7.1 Introduction

Biostatistics is the application of statistical methods to the solution of biological problems. Biostatistics can also defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine. *Statistics* is the study of collection, analysis, interpretation, presentation, and organization of data. Statistics is the scientific study of numerical data based on natural phenomenon. Scientific study deals with objectivity in presentation and evaluation data e.g. the arithmetic mean, standard deviations. Statistics is word also used to estimated value to describe a date.

7.2 objectives

- To know about basic of biostatistics
- To understand the biostatistics importance and applications in biology
- Practice the sampling techniques into biological problems

7.3 Scope

The field of Biostatistics is highly respected with right skills, can pursue an exciting and important career. For individuals with a strong interest in mathematics, science, statistics and health the Biostatistics is an excellent way to combine your strengths.

Biostatisticians are employed in health organizations, government agencies, universities, and even private companies. They are responsible for monitoring spread of disease and the causes of death to protect the population. Biostatistics is a billion dollar industry and it is an important part in improving human health and increasing knowledge of public health issues. The government and the pharmaceutical industry are the main contributors to the Biostatistics industry.

7.4 Applications in biology

Biostatistics covers applications and contributions not only from health, medicines and, nutrition but also from fields such as genetics, biology, epidemiology, and many others.

In physiology and anatomy

1. It's defined to find the normal or healthy population and to find limits of variability such as weight and pulse rate in population.
2. To find the correlation between two variables X&Y such as height and weight. In pharmacology to find the action of drugs.
3. To compare the actions of two different drugs or successive dosage of the same drug.
4. To find the relative potency of a new drug with respect to a standard drug.

In medicine

NOTES

1. To compare the efficacy of particular drug, operation or line of treatment. To find the attributes such as cancer and smoking and social class.
2. To identify the signs and symptoms of disease or syndrome. To test the usefulness of sera and vaccines in the field.
3. In clinical medicine used to document the medical history of the disease.
4. In preventive medicine it's used to provide the magnitude of any health problems in the community.
5. To introduce and promote health legislation.

Others

1. Used in the epidemiological studies
2. To find differences between means and proportions of normal at two places or in different periods

7.5 Sampling techniques

Sample is defined as a collection of individual observation from the population about which interferes is to be made and is obtained by a specific method. Thus a sample is a subgroup of the population.

The process of drawing a sample from a population is called sampling.

The sampling are based on two following principles

I. The law of statistical regularity.

II. The law of inertia of large numbers.

Sampling methods are the ways to choose people from the population to be considered in a sample survey. Samples can be divided based on following criteria.

NOTES

7.5.1 Random sampling

(1) Every element in our population has a nonzero probability of being selected as part of the sample.
(2) We have accurate knowledge of this probability, known as the inclusion probability, for each element in the sampling frame.
If both of these criteria are met, it is possible to obtain unbiased results about the population from studying the sample. To obtain unbiased results, it may sometimes be necessary to use weighting methods; such weighting is possible precisely because we know each individual's probability of being included in the sample. Samples obtained under these conditions are also known as **random samples.**

7.5.1.1 Simple random sample

Every member and set of members has an equal chance of being included in the sample. Technology, random number generators, or some other sort of chance process is needed to get a simple random sample.

Example: A teachers puts students' names in a hat and chooses without looking to get a sample of students.

NOTES

Why it's good: Random samples are usually fairly representative since they don't favor certain members

7.5.1.1 Stratified random sample

The population is first split into groups. The overall sample consists of some members from every group. The members from each group are chosen randomly.

Example: A student council surveys 100100100 students by getting random samples of 252525 freshmen, 252525 sophomores, 252525 juniors, and 252525 seniors.

Why it's good: A stratified sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

Cluster random sample: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.

7.5.2 NON-RANDOM SAMPLING

For these reasons—and to minimize costs—researchers often turn to other sampling methods, known as nonrandom sampling. When using these alternative methods, researchers generally select elements for the sample based on hypotheses about the population of interest, known as selection criteria. For example, if we're selecting our sample by stopping people on the street, attempting to stop an equal number of men and women (to coincide with the

presumed gender distribution in the population) would be a criterion of nonrandom sampling.

In these cases, since the selection of units for the sample isn't random, we shouldn't talk about error estimates. In other words, a nonrandom sample tells us about a population, but we don't know how precisely: we can't determine a margin of error or a confidence level.

These types of sampling methods include availability sampling, sequential sampling, quota sampling, and discretionary sampling and snowball sampling.

7.5.2.1 Convenience sampling

Convenience sampling (also known as availability sampling) is a specific type of non-probability sampling method that relies on data collection from population members who are conveniently available to participate in study. Face book polls or questions can be mentioned as a popular example for convenience sampling.

Convenience sampling is a type of sampling where the first available primary data source will be used for the research without additional requirements. In other words, this sampling method involves getting participants wherever you can find them and typically wherever is convenient. In convenience sampling no inclusion criteria identified prior to the selection of subjects. All subjects are invited to participate.

In business studies this method can be applied in order to gain initial primary data regarding specific issues such as perception of image of a particular brand or collecting opinions of perspective customers in relation to a new design of a product.

In its basic form, convenience sampling method can be applied by stopping random people on the street and asking questionnaire questions. 'Pepsi Challenge' marketing campaign can be referred to as a relevant example for this sampling method. 'Pepsi Challenge' is occasionally held in large shopping centers and other crowded locations and all members of population are invited to participate in the contest without any discrimination.

Convenience sampling technique may prove to be effective during exploration stage of the research area, and when conducting pilot data collection in order to identify and address shortcomings associated with questionnaire design.

Example: One of the most common examples of convenience sampling is using student volunteers as subjects for the research. Another example is using subjects that are selected from a clinic, a class or an institution that is easily accessible to the researcher. A more concrete example is choosing five people from a class or choosing the first five names from the list of patients.

In these examples, the researcher inadvertently excludes a great proportion of the population. A convenience sample is either a collection of subjects that are accessible or a self selection of

individuals willing to participate which is exemplified by your volunteers.

7.5.2.2 Sequential Sampling

Sequential sampling is a non-probabilistic sampling technique, initially developed as a tool for product quality control. The sample size, n, is not fixed in advanced, nor is the timeframe of data collection. The process begins, first, with the sampling of a single observation or a group of observations. These are then tested to see whether or not the null hypothesis can be rejected. If the null is not rejected, then another observation or group of observations is sampled and the test is run again. In this way the test continues until the researcher is confident in his or her results.

This technique can reduce sampling costs by reducing the number of observations needed. If a whole batch of light bulbs is defective, sequential sampling can allow us to learn this much more quickly and inexpensively than simple random sampling. However, it is *not* a random sample and has other issues with making statistical inference.

7.5.2.3 Discretionary sampling

Discretionary sampling is a non-random sampling process where a seller selects specific loans from their portfolio for review and also used to select 'representative' units from the population (or) to infer that a sample is 'representative' of the population.

7.5.2.4 Snowball sampling

NOTES

Snowball sampling is a non-probability sampling technique that is used by researchers to identify potential subjects in studies where subjects are hard to locate.

Researchers use this sampling method if the sample for the study is very rare or is limited to a very small subgroup of the population. This type of sampling technique works like chain referral. After observing the initial subject, the researcher asks for assistance from the subject to help identify people with a similar trait of interest.

The process of snowball sampling is much like asking your subjects to nominate another person with the same trait as your next subject. The researcher then observes the nominated subjects and continues in the same way until the obtaining sufficient number of subjects.

Example:

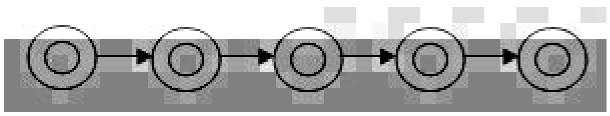
If obtaining subjects for a study that wants to observe a rare disease, the researcher may opt to use snowball sampling since it will be difficult to obtain subjects. It is also possible that the patients with the same disease have a support group; being able to observe one of the members as your initial subject will then lead you to more subjects for the study.

Types of Snowball Sampling

Linear Snowball Sampling

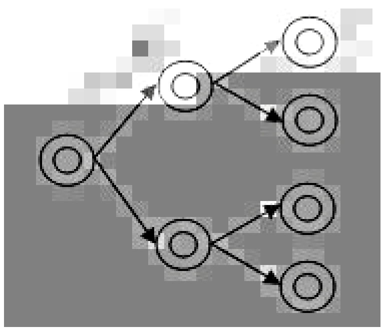
Formation of a sample group starts with only one subject and the subject provides only one referral. The referral is recruited into the

sample group and he/she also provides only one new referral. This pattern is continued until the sample group is fully formed.



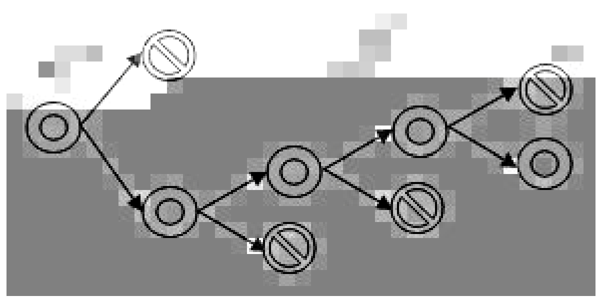
Exponential Non-Discriminative Snowball Sampling:

The first subject recruited to the sample group provides multiple referrals. Each new referral is explored until primary data from sufficient amount of samples are collected.



Exponential Discriminative Snowball Sampling:

Subjects give multiple referrals; however, only one new subject is recruited among them. The choice of a new subject is guided by the aim and objectives of the study.



NOTES

7.6 LET US SUM UP

In this unit, you have learnt about the basics of biostatistics and its importance to the biology. This knowledge would make understand what biostatistics is and how it plays major role biology. The sampling techniques have given detail knowledge about observation been occurred from the species. The various methods of sampling would pave the key attention to the student. This unit will be helpful to the student to apply the statistical importance into biology.

7.7 Unit-End Exercise

1. Define sampling techniques.
2. What are the major principles involved in the sampling?
3. Define Random sampling.

7.8 Check your progress

1. Define Biostatistics
2. Write down the applications of Biostatistics
3. Comment on the sampling methods

7.9 Answers to check your progress

1. Biostatistics is the application of statistical methods to the solution of biological problems. Biostatistics can also defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine.

2. Biostatistics covers applications and contributions not only from health, medicines and, nutrition but also from fields such as genetics, biology, epidemiology, and many others.
3. Two types of sampling methods were described such as random and non-random sampling methods.

7.10 SUGGESTED READINGS

1. Bernard Rosner (2006), Fundamentals of Biostatistics (6th Ed.), Thomson Brooks/Cole.
2. Wayne W. Daniel (2004), Biostatistics (9th Ed.), Wiley
3. Zar, J.H. (1984) "Bio Statistical Methods"; Prentice Hall International Edition, USA

NOTES

UNIT-VIII

Structure

NOTES

8.1 Introduction

Measures that describe the central aspects of a data are commonly referred to as averages. An average is a value that summarizes the characteristics of entire mass of the data. In samples, as well as in populations, one generally finds a preponderance of values somewhere in the middle of the range of observed values. Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution. As most of the items of the series are clustered around the average it is called a measure of central tendency. A measure of central tendency is also called a type as it is a typical value of the series. Since an average locates a distribution at some value of the variable, it is sometimes known as measure of location.

8.2. Objectives

1. The main aim of an average is to present huge mass of statistical data in a simple and concise manner
2. The averages are extremely useful for purpose of comparison.
3. Practice and learn mean, median and mode problems.

8.3 Mean

NOTES

The arithmetic mean is the sum of the individual values in a data set divided by the number of values in the data set. We can compute a mean of both a finite population and a sample. For the mean of a finite population (denoted by the symbol), we sum the individual observations in the entire population and divide by the population size, N. When data are based on a sample, to calculate the sample mean (denoted by the symbol (–X) we sum the individual observations in the sample and divide by the number of elements in the sample, n. The sample mean is the sample analog to the mean of a finite population.

$$\bar{x} = \frac{\sum_{I=1}^{n} x\,i}{N}$$

$\overline{X}$ = Mean, $\sum X$ = Sum of all the items, i = Variants of from 1 to n, N = Total no of items

The population mean (and also the population variance and standard deviation) is a parameter of a distribution. Means, variances, and standard deviations of finite populations are almost identical to their sample analogs. We will learn more about these terms and appreciate their meaning for infinite populations after we cover absolutely continuous distributions and random variables. Statisticians generally use the arithmetic mean as a measure of central tendency for numbers that are from a ratio scale (e.g., many biological values, height, blood sugar, cholesterol), from an

interval scale (e.g., Fahrenheit temperature or personality measures such as depression), or from an ordinal scale (high, medium, low). The values may be either discrete or continuous; for example, ranking on an attitude scale (discrete values) or blood cholesterol measurements (continuous). It is important to distinguish between a continuous scale such as blood cholesterol and cholesterol measurements. While the scale is continuous, the measurements we record are discrete values. For example, when we record a cholesterol measurement of 200, we have converted a continuous variable into a discrete measurement.

Example:

Suppose we have measured the length in *cm* of 8 snakes. The ungrouped data reads as follows,

In the following example, length is the variable *X*, *n*=8

2.5 3.4 4.3 3.3 4.2 2.9 3.0 3.1

$$\bar{x} = \frac{\sum_{I=1}^{n} x}{n}$$

$$\bar{x} = x_1 + x_2 + x_3 + x_4 + x_5 + x_6/n$$

$$\bar{x} = \frac{2.5+3.4+4.3+3.3+4.2+2.9+3.0+3.1}{8}$$

$$\bar{x} = \frac{26.7}{8} = 3.375$$

NOTES

8.4 Median

Median is a central value of the distribution, or the value which divides the distribution in equal parts, each part containing equal number of items. Thus it is the central value of the variable, when the values are arranged in order of magnitude.

Connor has defined as "The median is that value of the variable which divides the group into two equal parts, one part comprising of all values greater, and the other, all values less than median" Median is the value of the middle item of a given series of data arranged in ascending order of magnitude, i.e., in an array.

It divides the array into two equal parts. The two parts on the either side of the median have the following properties:

1. The numbers of items on both sides of the median are equal.

2. The values of items in front of the median are lesser than the values after the median.

Median = value of the item (n+1)/2 in an array

The median should be expressed in the same unit as that of the data.

Example: The following are the length in cm of a species of snakes. Let us identify the median weight.

NOTES

Length in cm: 17, 16,15,18,16

First let us array the data in ascending order of magnitude 15, 16, 16, 17, 18

Then, we identify the median as the value of the item (n+1)/2, where n is the number of sample. It is 5 in this example.

Median = value of item (n+1)/2

= (5+1)/2 = 6/2

= 3.

That is, the median is the value of the item 3 in the array. In the above example the value of the item 3 is 16.

The median length of the sample is 16 cm

8.4.1 Determining a Median from a Frequency

Class Interval	*f*	*cf*
160-169	1	100
150-159	3	99
140-149	1	96
130-139	2	95
120-129	6	93
110-119	4	87
100-109	7	83
90-99	37	76
80-89	33	39
70-79	6	6

NOTES

Previously, we stated that the measurements from a continuous scale represent discrete values. The numbers placed in the frequency table were continuous numbers rounded off to the nearest unit. The real limits of the class interval are halfway between adjacent intervals. As a result, the real limits of a class interval, e.g., 90–99, are 89.5 to 99.5. The width of the interval (i) is (99.5 – 89.5), or 10. Thus, placing these values in Formula yields,

$$\textbf{Median} = 89.5 + 10[(0.50)(100) - 39] = 97.47$$

For data that have not been grouped, the sample median also can be calculated in a reasonable amount of time on a computer. The computer orders the observations from smallest to largest and finds the middle value for the median if the sample size is odd. For an even number of observations, the sample does not have a middle value; by convention, the sample median is defined as the average of two values that fall in the middle of a distribution. The first number in the average is the largest observation below the halfway point and the second is the smallest observation above the halfway point.

Let us illustrate this definition of the median with small data sets. Although the definition applies equally to a finite population, assume we have selected a small sample. For n = 7, the data are {2.2, 1.7, 4.5, 6.2, 1.8, 5.5, 3.3}. Ordering the data from smallest to largest, we obtain {1.7, 1.8, 2.2, 3.3, 4.5, 5.5, and 6.2}. The middle observation (median) is the fourth number in the sequence; three

values fall below 3.3 and three values fall above 3.3. In this case, the median is 3.3.

NOTES

Suppose n = 8 (the previous data set plus one more observation, 5.7). The new data set becomes {1.7, 1.8, 2.2, 3.3, 4.5, 5.5, 5.7, and 6.2}. When n is even, we take the average of the two middle numbers in the data set, e.g., 3.3 and 4.5. In our example, the sample median is (3.3 + 4.5)/2 = 3.9. Note that there are three observations above and three below the two middle observations.

8.5 The Mode

The mode refers to the class (or midpoint of the class) that contains the highest frequency of cases. The modal class is 90–99. When a distribution is portrayed graphically, the mode is the peak in the graph. Many distributions are multimodal, referring to the fact that they may have two or more peaks. Such multimodal distributions are of interest to epidemiologists because they may indicate different causal mechanisms for biological phenomena, for example, bimodal distributions in the age of onset of diseases such as tuberculosis, Hodgkins disease, and meningococcal disease. The mode illustrates unimodal and bimodal distributions.

Example: Weight of snake in g: 8,10,9,17,10,19,15,10,12,19.

Array of the data: 8, 9, 10, 10, 10, 12, 15,17,19,19.

In the above data, 10 occurs 3 times,

19 occurs 2 times,

NOTES

Others each one time

Therefore, the mode of the above data is 10g.

8.6 Variance (S2)

Variance (S2) = average squared deviation of values from mean

Calculating variance involves squaring deviations, so it does not have the same unit of measurement as the original observations. For example, lengths measured in meters (m) have a variance measured in meters squared (m2).

Taking the square root of the variance gives us the units used in the original scale and this is the standard deviation.

8.7 Standard deviation (S)

Standard deviation (S) = square root of the variance

Standard deviation is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency. Thus, it measures spread around the mean. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency.

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation. In that sense, the standard deviation is a good indicator of the presence of

outliers. This makes standard deviation a very useful measure of spread for symmetrical distributions with no outliers.

Standard deviation is also useful when comparing the spread of two separate data sets that have approximately the same mean. The data set with the smaller standard deviation has a narrower spread of measurements around the mean and therefore usually has comparatively fewer high or low values. An item selected at random from a data set whose standard deviation is low has a better chance of being close to the mean than an item from a data set whose standard deviation is higher.

Generally, the more widely spread the values are, the larger the standard deviation is. For example, imagine that we have to separate two different sets of exam results from a class of 30 students the first exam has marks ranging from 31% to 98%, the other ranges from 82% to 93%. Given these ranges, the standard deviation would be larger for the results of the first exam.

Standard deviation might be difficult to interpret in terms of how big it has to be in order to consider the data widely spread. The size of the mean value of the data set depends on the size of the standard deviation. When you are measuring something that is in the millions, having measures that are "close" to the mean value does not have the same meaning as when you are measuring the weight of two individuals. For example, a measure of two large companies with a difference of $10,000 in annual revenues is considered pretty close, while the measure of two individuals with

NOTES

a weight difference of 30 kilograms is considered far apart. This is why, in most situations, it is useful to assess the size of the standard deviation relative to the mean of the data set.

Although standard deviation is less susceptible to extreme values than the range, standard deviation is still more sensitive than the semi-quartile range. If the possibility of high values (outliers) presents itself, then the standard deviation should be supplemented by the semi-quartile range.

8.7.1 Properties of standard deviation

When using standard deviation keep in mind the following properties.

Standard deviation is only used to measure spread or dispersion around the mean of a data set.

Standard deviation is never negative.

Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread.

For data with approximately the same mean, the greater the spread, the greater the standard deviation

If all values of a data set are the same, the standard deviation is zero (because each value is equal to the mean).

When analyzing normally distributed data, standard deviation can be used in conjunction with the mean in order to calculate data intervals.

If x bar = mean, S = standard deviation and x = a value in the data set, then

- About 68% of the data lie in the interval: mean - S < x < mean + S.
- About 95% of the data lie in the interval: mean - 2S < x <mean + 2S.
- About 99% of the data lie in the interval: mean - 3S < x < mean + 3S.

8.7.2 Discrete variables

The standard deviation for a discrete variable made up of n observations is the positive square root of the variance and is defined as:

$$S = \sqrt{\sum_{i=1}^{n}(Xi - \overline{X})^2}$$

Use this step-by-step approach to find the standard deviation for a discrete variable.

1. Calculate the mean.
2. Subtract the mean from each observation.
3. Square each of the resulting observations.

NOTES

4. Add these squared results together.
5. Divide this total by the number of observations (variance, S2).
6. Use the positive square root (standard deviation, S).

8.7.3 Example of standard deviation

A hen lays eight eggs. Each egg was weighted and recorded as follows;

60g, 56g, 61g, 68g, 51g, 53g, 69g, 54g.

a. First, calculate the mean.

$\overline{X} = \sum X/n$

$= 472/8 = 59$

B. now, fined the standard deviation.

Weight of eggs, in grams

Weight (x)	(x - mean)	(x - mean)2
60	1	1
56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25

472		**320**

In order to calculate the standard deviation, we must use the following formula:

SD = $\sqrt{(X - \bar{X}/n}$

SD = $\sqrt{320/8}$

= 6.32 grams

8.8 Standard error

The standard error (SE) is very similar to standard deviation. Both are measures of spread. The higher the number, the more spread out data is. To put it simply, the two terms are essentially equal — but there is one important difference. While the standard error uses statistics (sample data) standard deviations use parameters (population data).

In statistics, you'll come across terms like "the standard error of the mean" or "the standard error of the median." The SE tells how far sample statistic (like the sample mean) deviates from the actual population mean. The larger sample size the smaller the SE. In other words, the larger sample size, the closer sample mean is to the actual population mean.

8.8.1 The SE Calculation

The calculation is different for the mean or proportion. We asked to find the sample error; we're probably finding the standard error.

NOTES

That uses the following formula: s/√n. we might be asked to find standard errors for other stats like the mean or proportion.

8.8.2 The Standard Error Formula:

The following tables show how to find the Standard Error. That assumes we know the right population parameters. If we don't know the population parameters, we can find the standard error:

- Sample mean.
- Sample proportion.
- Difference between means.
- Difference between proportions.

Statistic (Sample)	Formula for standard error.
Sample mean X	= s / sqrt (n)
Sample proportion, p	= sqrt [p (1-p) / n)
Difference between means x1-x2	= sqrt [$s^2_1/n_1 + s^2_2/n_2$]
Difference between proportions p1.p2	= sqrt [$p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$]

8.9 Let us sum up

In this we focused the measures of central tendency and standard deviation. This unit will be able to provide basic statistics an important indicator that the amount of effort students spends on

learning statistics do not necessarily warrant a good understanding of the measures of central tendency. Specifically, it might be useful to introduce the appropriate learning technique in which students can realize that they can master the topic and develop better understanding in this discipline.

8.10 Unit-End Exercise

1. Define mean, median and mode.
2. Give an account on Properties of standard deviation?
3. Define Discrete variables with formula?
4. Why Standard error occurs?

8.11 Check your progress

1. How do you calculate the mean?
2. What is a high standard deviation?
3. . What factors affect standard error?

8.12 Answers to check your progress

1. The mean is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are. In other words it is the sum divided by the count.

2. A low standard deviation indicates that the data points tend to be very close to the mean. A high standard deviation indicates that the data points are spread out over a large range of values.

3. The major factor affects standard error of the mean is sample size. The size of the sample increases the standard error of the

NOTES

mean decreases. Another factor affecting the standard error of the mean is the size of the population standard deviation.

8.13 SUGGESTED READINGS

1. Gurumani,N., (2015).“An Introduction to Biostatistics”, MJP Publisher, 2nd Edition

2. Warren, J., Gregory, E. and Grant, R. (2004) “Statistical methods in Bioinformatics”; First edition, Springer-Verlag, Berlin.

3. Milton, J.S. (1992) “Statistical methods in the Biological and Health Sciences”; Second Edition, McGraw Hill Publishers

UNIT IX

9.1 Introduction

The term probability is commonly used one, to mean “chance” or “like hood” of occurrence of an event. We say “it will probably rain today”. What we mean is that we “expect” it to rain. This expectation of the event of rain might be based on our experience such as the sultry conditions, high day temperature or cloudy

NOTES

conditions etc. It may even indicate our desire for an event to occur. In statistics, the term probability is used in a more precise sense. It is expressed numerically. The symbol to denote probability is P. The probability is measured in a probability scale. The probability scale runs from 0 to 1. There is no probability lesser than 0, i.e., there is no negative probability; and there is no probability greater than unity, i.e., 1.

In nature, events that have a probability of exactly 1 or 0 are rare. Most often events have probability values between 0 and 1.

9.2 objectives

- Define event, outcome, trial, simple event, sample space and calculate the probability that an event will occur.
- Calculate the probability of events for more complex outcomes.
- Solve applications involving probabilities.

9.3 Terminology

We shall define a few terms that are commonly used in measurements of probability and explain them with common examples such as tossing a coin or rolling of dice (a cube with sides marked 1 to 6).

9.4 Event

Each possible outcome of experiments is called an event.

Example: 1

NOTES

When you toss a coin the result would be either” Head” or “Tail”. Tossing of the coin is the experiments. The “Head” and the “Tail” are the possible outcomes. The “Head” is an event. The “Tail” is an event.

Example: 2

When role a dice there is possible event of getting 1, 2,3,4,5 and 6

Example: 3

When you measure the length of a fish, the results of the measurements is an event. Suppose you measure the length of the fish as 12.5 cm. This length 12.5 cm is an event.

9.5 Exhaustive event

A group of events is said to be exhaustive if it includes all possible events.

Example

In the tossing of the coin experiments, the possible events are the “Head” and the “Tail”. No other event is possible. Therefore, the group of events “Head and Tail” is exhaustive events. Similarly the events 1 to 6 are the exhaustive events in the experiments of rolling the dice.

9.6 Equiprobable events

A group of events is said to be equiprobable events if there is equal chance for each event in the group to occur.

Example:

The event “Head and Tail” in the coin tossing experiment are equiprobable events. When a coin is tossed there is no reason to expect anyone to occur rather than the other. Similarly the events 1 to 16 in the dice rolling experiments are equiprobable events.

9.7 Mutually exclusive events

Two or more events are said to be mutually exclusive if the occurrence of one event excludes the occurrence of the other.

Example

“Head and Tail” in coin tossing are mutually exclusive events. When you toss a coin, occurrence of “Head” will exclude the occurrence of the “Tail”, and vice versa. In the dice rolling, the events 1 to 6 are mutually exclusive, because the occurrence of any one number on top, would exclude the occurrence of other numbers.

9.8 Independent events

Two or more events said to be independent events if the occurrence of any one of them does not in any way affect the occurrence of the other.

Example:

Suppose you are tossing the coins two times. The outcome of each toss is independent of other. You may get Head in the first toss. This event is in no way going to affect what you are going to get in

the second toss. Similarly, you can roll the dice several times. The result of each roll is independent of the results of other rolls.

9.9 Kind of probability

The probability majorly occurs in two types for the purpose computation of the value of probability of occurrence of an event.

I. Apriori probability(mathematical probability)
II. Aposteriori probability(statistical probability)

I Apriori probability

Here the probability is computed based on established facts.

Computed of Apriori probability:

If an event can happen in m ways and fail to happen in n ways and all these ways are mutually exclusive and equiprobable, then the probability of the occurrence of the event is m/ (m+n). It can also be described as

$$\textbf{Apriori probability (P)} = \frac{\textbf{Number of favorable cases}}{\textbf{Total number of possible cases}}$$

The event H can occur in 1 way. The event H can fail to occur in 1 way, i.e., if you get T. Therefore probability of the occurrence of the event H, P (H) =1/ (1+1) =1/2. It can be also calculated as,

$$\textbf{P (H)} = \frac{\textbf{Number of favorable cases}}{\textbf{Total number of possible cases}}$$

$$\textbf{P (H)} = \frac{\mathbf{1}}{\mathbf{2}} = 1/2$$

NOTES

II Aposteriori probability

Aposteriori probability (statistical or empirical probability) is a ratio of the number of occurrences of an event to the total number of trials.

Computation of Aposteriori probability:

$$P = \frac{\textbf{Number of times the event occurred}}{\textbf{Total number of trials}}$$

The relative frequencies of a frequency distribution are in fact the Aposteriori probabilities of the occurrence of the respective classes or class-intervals.

Example

Simple and relative frequencies number of children/ family in 350 families of a city.

No of children/family	No. of. families	Relative frequencies=P
O	20	20/350
1	80	80/350
2	160	160/350
3	70	70/350
4	10	10/350
5	7	7350
6	3	3/350

Total	350	1.000

From these we can find the probability of a randomly chose family of this city having 3 children's is 0.200, i.e., the number of times the event (three-children family) occurred is 70 and the total number of trials is 350 and the ratio between these two is 0.200.

9.10 Binomial distribution

Binomial distribution describes the distribution of discrete events that can be classified in a dichotomous (binomial) way of "occurrence of the event" and "non-occurrence of the event" in a specific number of trials or samples (n).

Example1: The event "white eye" occurs in Drosophila or fails to occur. "Fails to occur" here means the eye-color may be anything (red, blue etc.) but not white. Number of white-eyed flies can be counted in a sample of 10 flies.

- Assume an experiment consisting of *n* trials is conducted, and the probability of occurrence of a particular event is p in each trial and the probability of the non-occurrence of the same event is q [= 1-p] in each trial.
- Now the probability of getting exactly $n, n\text{-}1, n\text{-}2, \ldots, n\text{-}r, \ldots,$ 2, 1, 0 occurrences (probability distribution) are given by the successive terms in the expansion of $(p+q)^n$. The probability distribution so obtained is called the "binomial distribution".

NOTES

- In the above experiment consisting of n trials is repeated N times, then the number of experiments (frequency) in which the event occurs *n*, *n-1*, *n*-2,…., *n*-r,…., 2, 1, 0 times is given by the successive terms of the expansion of *N* $(p+q)^n$.

9.11 Properties of Binomial distribution

1. The general form (shape) of the binomial distribution depends on the values of its parameters, p, q and n. If p and q are equal, the given binomial distribution will be symmetrical. If p and q are not equal, the distribution will be skewed distribution.
2. Even though p and q are not equal, if n is very large, the distribution will tend to be a symmetrical one.

Binomial equation

$$(p+q)^n = p^n + {}_nC_{n-1}p^{n-1}q + {}_nC_{n-2}p^{n-2}q^2 + ... + {}_nC_r p^r q^{n-r} ... + {}_nC_1 pq^{n-1} + q^n$$

No. of times the event occurs	Probability
n	p^n
$n-1$	${}_nC_{n-1}p^{n-1}q$
$n-2$	${}_nC_{n-2}p^{n-2}q^2$
$n-r$	${}_nC_r p^r q^{n-r}$
2	${}_nC_2 p^2 q^{n-2}$
1	${}_nC_1 pq^{n-1}$
0	q^n

NOTES

Example: What is the probability of occurrence of 6 Heads in 10 tosses?

Step1: n=10; r=6; p=P (H) =1/2; q=P (H) = 1-p = 1-1/2= 1/2.

Step2: P (r events in n trials) = ${}_{n}C_{r}p^{r}q^{n-r}$

P (6 H in 10 tosses) = ${}_{10}C_{6}p^{6}q^{4}$

(${}_{10}C_{6}$ = 10p6/6! = (10x9x8x7x6x5)/(6x5x4x3x2x1) = 210)

= 210 $(1/2)^{6}(1/2)^{4}$ = 210 x 1/64 x 1/16

= 210/1024 = 0.205

Probability of getting 6 Heads in 10 tosses is 0.205:

9.12 Poisson distribution

- It is a distribution of the number of times a rare discrete event occurs in a continuum of space or time.
- In Poisson distribution, n is not known or it is infinitely large and, therefore, you can count only the number of times the rare event occurs but not the number of times the event fails to occur.
- Poisson distribution is generally studied in a spatial or temporal sample.

Example: A 100 km stretch of trunk road through a forest was surveyed to determine the incidence of death of wild life due to accidents caused by heavy vehicles. The total number of dead

NOTES

animals counted was 75. Calculate the probability of finding no dead wild life in any randomly selected km stretch of this trunk road. What is the probability of finding one dead animal? Two died animals?

- **Step1:** Occurrence of dead wild life per km stretch of trunk road is a Poisson event; because it is a rare, discrete event in a continuum of space, i.e., stretch of trunk road.
- **Step2:** Poisson parameter m = mean number of dead wild life per km stretch of trunk road
  - o m = 75/100 = 0.75
- **Step3:** If m = 0.75, $e^{-m} = e^{-0.75} = 0.472$
- **Step4:** P(x event) = $e^{-m}m^{x}/x!$
  - P (0 dead animal) = $e^{-m}m^{0}/0! = 0.472$ x 0.750/0! = 0.472 x1/1 = 0.472
    - o (Any number raised to the power of 0 = 1 and 0! = 1)
  - P (1 dead animal) = $e^{-m}m^{1}/1! = 0.472$ x 0.75/1 = 0.354
  - P (2 dead animal) = $e^{-m}m^{2}/2! = 0.472$ x $0.75^{2}/2! = 0.133$

9.13 Normal Distribution

- The mathematical model of continuous variable is given by normal distribution.

NOTES

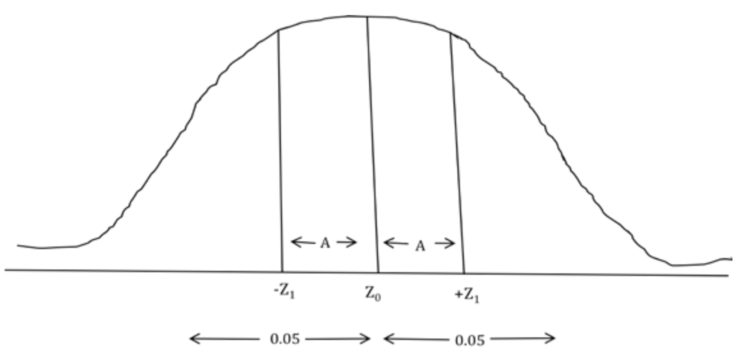
- Normal distribution is derived by its graphical expression called normal curve (also called as Gaussian curve, Laplacian curve).
- The normal curve is a typical bell – shaped symmetrical curve. The mathematical equation that gives a height of the curve at any specific value is as follows:

$$Y = \frac{N}{6\sqrt{2\pi}}.e^{\frac{-(X-\overline{X})^2}{26^2}}$$

Y= Ordinate of the point at a distance x from the origin, i.e., the probability density of a particular value of the variable X.

N = Total frequency which is the total area under the curve.

σ = standard deviation; π = 3.14159; e = 2.7183; = a particular value of variable = Mean of distribution.



Probability of the value	Formula
1. Lying between 0 and + Z	A
2. Lying between - Z and + Z	2A(A + A)
3. Lying outside the interval (-Z, Z)	1 - 2A
4. Less than Z (Z positive)	0.5 + A
5. Less than Z (Z negative)	0.5 - A
6. Greater than Z (Z positive)	0.5 - A
7. Greater than Z (Z negative)	0.5 + A
8. Lying between two Zs (both + and -)	Larger A - Smaller A

NOTES

Example: The mean and SD of a Normal distribution are respectively 20 and 10. Find out the probability that x will lie between 25 and 35.

Step1: = 20; SD = 10; = 25 – 35

Step2: calculation of Z = (x-)/SD

When x = 25, Z = (25-20)/10 = 5/10 = 0.5

When x = 35, Z = (35-20)/10 = 15/10 = 1.5

Step3: Area between Z_0 and $Z_{0.5}$ = 0.1915

Area between Z_0 and $Z_{1.5}$ = 0.4332

Area between $Z_{0.5}$ and $Z_{1.5}$ = 0.4332 – 0.1915 = 0.2417

Step4: P(x lies between 25 and 35) = 0.2417.

9.14 Let us sum up

Probability helps students to overcome their deterministic thinking and accept the existence of chance in nature. In this effort, however, students and teachers are faced with some tensions and challenges such as being able to conceptualize the connection between the natural and academic language of probability.

9.15 Unit-End Exercise

Define the term probability.

What is an event and why it's occurred?

Define mutually exclusive events and suggest with example?

.

NOTES

9.16 Check your progress

1. What is probability of mutually exclusive event?
2. What are the properties of a binomial distribution?
3. What is normal distribution Z?

9.17 Answers to check your progress:

1. Two events are mutually exclusive if they cannot occur at the same time. Another word that means mutually exclusive is disjoint. ... If two events are mutually exclusive, then the probability of either occurring is the sum of the probabilities of each occurring.

2. Binomial distribution is applicable when the trials are independent and each trial has just two outcomes success and failure. It is applied in coin tossing experiments, sampling inspection plan, genetic experiments and so on.

3. The standard normal distribution table provides the probability that a normally distributed random variable Z, with mean equal to 0 and variance equal to 1, is less than or equal to z.

9.18 SUGGESTED READINGS

1. Seymour Lipschutz and John Schiller (1998), Schaum's Outlines - Introduction to Probability and Statistics, TATA McGraw-Hill edition.

2. Rosner, B. (2005) "Fundamentals of Biostatistics"; Duxbury Press.

3. Wayne W. Daniel (2004), Biostatistics (9th Ed.), Wiley

UNIT X

NOTES

10.1 Introduction

The Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical

variable match the expected frequencies for the categorical variable. The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.

We try to test the likelihood of test data (sample data) to find out whether the observed distribution of data set is a statistical fluke (due to chance) or not. “Goodness of fit” statistic in chi-square test, measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

10.2 objectives

- Perform chi-square tests by hand
- Appropriately interpret results of chi-square tests
- Identify the appropriate hypothesis testing procedure based on type of outcome variable and number of samples

10.3 Type of Chi-Square Test:

10.3.1. For One Categorical Variable, we perform, **Chi-Square Goodness-of-Fit Test**

The chi square goodness of fit test begins by hypothesizing that the distribution of a variable behaves in a particular manner. For example, in order to determine daily staffing needs of a retail store, the manager may wish to know whether there are an equal number of customers each day of the week.

NOTES

10.3.2. For Two Categorical Variables, we perform, **Chi-Square Test for Association**

The test compares the observed data to a model that distributes the data according to the expectation that the variables are independent. Wherever the observed data doesn't fit the model, the likelihood that the variables are dependent becomes stronger, thus proving the null hypothesis incorrect!

10.4 Chi-Square Statistic:

The formula for the chi-square statistic used in the chi square test is:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

The subscript "c" here is the degrees of freedom. "O" is your observed value and E is your expected value. The summation symbol means that you'll have to perform a calculation for every single data item in your data set.

E= (row total × column total) / sample size

The Chi-square statistic can only be used on the numbers. They can't be used for percentages, proportions, means or similar statistical value. For example, if you have 10 percent of 200 people, you would need to convert that to a number (20) before you can run a test statistic.

10.5 P-Value:

NOTES

The null hypothesis provides a probability framework against which to compare our data. Specifically, through the proposed statistical model, the null hypothesis can be represented by a probability distribution called P-value, which gives the probability of all possible outcomes if the null hypothesis is true;

p2+ 2pq + q2 = 1

PP – Dominant allele.

Pp - Recessive allele.

pp - Mutant allele.

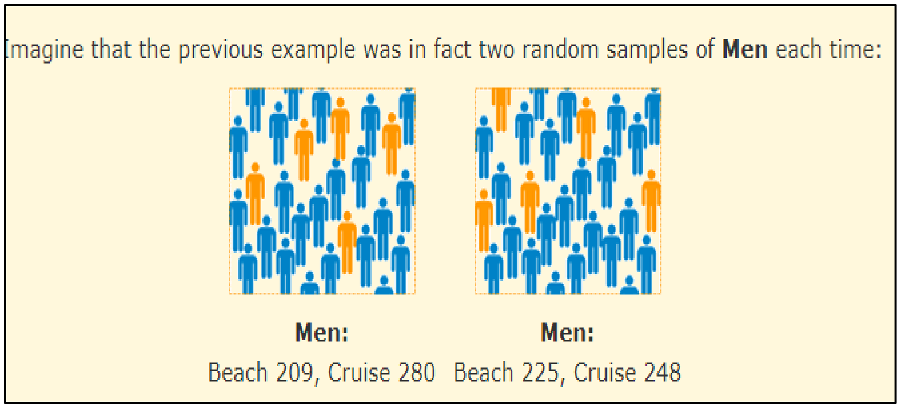
10.6 Example of Chi-square test;

Taken two groups and put them in categories single, married or divorced;

The Chi-Square Test gives a "p" value to help you decide! Because people want to see a p value less than **0.05** before they are happy to say the results show the groups have a different response.

"p" is the probability the variables are independent.

NOTES



10.7 Important points before we get started:

This test only works for categorical data (data in categories), such as Gender {Men, Women} or color {Red, Yellow, Green, Blue} etc, but not numerical data such as height or weight.

The numbers must be large enough. Each entry must be 5 or more. In our example we have values such as 209, 282, etc, so we are good to go.

The two hypotheses are.

Gender and preference for cats or dogs are independent.

Gender and preference for cats or dogs are not independent.

Lay the data out in a table:

	Cat	Dog
Men	207	282
Women	231	242

Add up rows and columns:

	Cat	Dog	
Men	207	282	489
Women	231	242	473
	438	524	962

NOTES

Calculate "Expected Value" for each entry:

Multiply each row total by each column total and divide by the overall total:

	Cat	Dog	
Men	$\frac{489x438}{962}$	$\frac{489x524}{962}$	489
Women	$\frac{473x438}{962}$	$\frac{473x524}{962}$	473
	438	524	962

Gives us

	Cat	Dog	
Men	222.64	266.36	489
Women	214.36	257.64	473
	438	524	962

Subtract expected from actual, square it, then divide by expected:

NOTES

	Cat	Dog	
Men	$\frac{(207-222.64)^2}{222.64}$	$\frac{(282-266.36)^2}{266.36}$	489
Women	$\frac{(231-214.36)^2}{214.36}$	$\frac{(242-257.64)^2}{257.64}$	473
	438	524	962

Which is,

	Cat	Dog	
Men	1.099	0.918	489
Women	1.136	0.949	473
	438	524	962

Now add up those values:

1.099+0.918+1.136+0.949 = 4.102

Chi-square is **4.102**

Calculate Degrees of Freedom

Multiply (rows − 1) by (columns − 1)

Example: DF = (2 − 1) (2 − 1) = 1×1 = 1

Result:

The result is: **p = 0.04283**, so the chi-square was done.

NOTES

10.8 Goodness-of-fit test and null hypothesis

Example: A coin was tossed 100 times. Results were counted. 65 heads and 35 tails were obtained. Is this result a significant departure from the expected result?

Step1: Calculation of expected frequencies

Total number of trials (n) = 100

Observed frequencies (O): Head (H) - 65; Tail (T) – 35;

- Expected frequencies (E) to be calculated based on apriori hypothesis.

We know that in coin tossing experiments P (H) = ½ and P (T) = ½.

Therefore, the expected frequency of H = P(H)×n = ½ × 100 = 50.

The expected frequency of T = P (T)×n = ½ × 100 = 50.

- Step 2: Presentation of the results in 2×2 table

	Heads	**Tails**	**Row total**
Observed	65	35	100

NOTES

Expected	50	50	100
Column total	110	85	200

- Step 3: Degree of Freedom (DF) = (rows-1)(columns-1) = (r-1)(c-1)

= (2-1) (2-1) = 1

- Step 4: Null-hypothesis: H_o: O-E = 0

(The hypothesis in this problem is that the observed frequencies of Head and Tail, viz., 65 H and 35 T, is a significant departure from the expected 50 H and 50 T.)

- When there is no significant difference between the O and E frequencies, symbolically it is stated H_o: O-E = 0.

Step 5: Level of significance: 0.05

Step 6: Computation of $\chi 2$ (with Yate's correction factor)

$$X^2 = \sum \frac{[(O-E)-0.5]^2}{E}$$

(Yate's correction factor is the subtraction of 0.5 from the absolute difference between O and E combination. This is to be used only with 2×2 table.)

$$X^2 = \frac{[(65-50)-0.5]^2}{50} + \frac{[(35-50)-0.5]^2}{50}$$

NOTES

$$= \frac{(14.5)^2}{50} + \frac{(14.5)^2}{50}$$

$$= \frac{210.25}{50} + \frac{210.25}{50}$$

$$= 4.2 + 4.2 = 8.4$$

8.4 is the calculated X^2 value.

- **Step 7:** Decision about the Ho (reject H_o or fail to reject it)

Enter the χ^2 table with 1DF and 0.05 LS. You find a value of 3.841. This is table χ^2 value. Compare your calculated χ^2 value (8.4) with the table χ^2 value (3.841)

The decision is made as follows:

If the calculated χ^2 value is greater than table χ^2 value, reject H_o.

In our example the calculated χ^2 value is 8.4 is certainly greater than table χ^2 value of 3.841. Therefore, we reject the H_o.

We have reject it because the probability of its occurrence is less than 0.05 ($P<0.05$)

- **Sep8:** Inference:

We have rejected the H_o: O-E = 0. i.e., there is no significant difference between the observed and expected frequencies. It means, there is significant difference between the observed and expected frequencies.

NOTE

Result: Yes, this result is a significant departure from the expected result.

10.9 Let us sum up

Chi-square is used most commonly to compare the incidence (or proportion) of a characteristic in one group to the incidence (or proportion) of a characteristic in other group(s). Students can be used to evaluate whether there is an association between the rows and columns in a contingency table. In this unit may helpful to the student to solve their own biological problems using Chi-square test.

10.10 Unit-End Exercise

1. Comment on the type of Chi-Square Test.
2. Define Chi-Square Statistic.
3. What is P-Value?
4. Give an Example that satisfies the Chi-square test;

10.11 Check your progress

1. What is the Type involved in Chi-Square Test?
2. Why is p value important?
3. How do you accept or reject the null hypothesis in Chi Square?

10.12 Answers to check your progress

1. Types of chi-square test

For One Categorical Variable, we perform, **Chi-Square Goodness-of-Fit Test**

For Two Categorical Variables, we perform, **Chi-Square Test for Association**

2. The p-value is the probability that the null hypothesis is true. ... A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance. A non-significant result, leading us not to reject the null hypothesis, is evidence that the null hypothesis is true.

3. If your chi-square calculated value is greater than the chi-square critical value, then you reject your null hypothesis. If your chi-square calculated value is less than the chi-square critical value, then you "fail to reject" your null hypothesis.

10.13 SUGGESTED READINGS

1. Zar, J.H. (1984), Bio Statistical Methods, Prentice Hall International Edition, USA

2. Gurumani,N., (2015), An Introduction to Biostatistics (2nd Ed.), MJP Publisher,

3. Dutta, N. K. (2004). Fundamentals of Biostatistics, Kanishka Publishers.

NOTES

BLOCK 4: Statistical Analysis

UNIT XI

1.1 Introduction

The "Analysis of variance" (ANOVA) is the appropriate statistical technique to be used in situations where we have to compare more than two groups. In the unit we considered under hypothesis-testing instances involving comparison of two groups using statistics from either large samples. In biology, we come across many situations where in we have to compare the parameters of more than two populations. Use of F-test is rather difficult in such situations for the obvious reason of the increasing number of tests we will have to perform to compare each pair of means in all possible combinations. Further, the probability of error would become enormous, proportional to number of tests. The "Analysis of

variance” (ANOVA) is the appropriate statistical technique to be used in situations where we have to compare more than two groups.

NOTES

11.2 Objectives

- To understand Analysis of variance and its importance
- To perform F-test anvd their variance
- Practice one way and two way classifications

11.3 Methods of ANOVA

11.3.1 One way classification Design

This is relatively a simple and common design in which two or more groups are compared. More specifically this design is described as one way classification, completely random design with fixed effects. It is referred to as one way classification because only one factor is under study in each experiment. The factor may be studied at different levels, i.e., in a number of groups. In other words, one factor is studied in k groups.

Let us consider an example to illustrate the above. Suppose we are interested in studying the effect of temperature on bacterial growth. In such study the factor is the growth of bacteria measured in terms of number of colonies/Plate. The different temperatures such as 25º, 30º,35º,40º,and 45º at which the bacteria are cultured are the levels, i.e., 5 groups (k=5). At each temperature we may have a number of replicates, i.e., culture plates, which refers to the sample size in each level or group. The number of replicates in different groups may be equal or unequal. The samples of the k groups are independent of on another. The allotment of a culture plate is random. Hence the design is completely random design. On the other hand we decide the number of

NOTES

levels. Though we could have selected any number of temperatures, say, from 0 º to 100 º, we chose only 5 specific temperatures. Even these temperatures were not selected randomly. That is why this design of experiment is described to have fixed effects, meaning that the levels of factor are specifically selected by the investigator because of their significance.

In the above experimental model, the factor, growth of bacteria (number of colonies/plates), may be studied in a number of bacterial species at a particular temperature, say, 37º. Here the levels of the factor are the number of species. Each species is cultured in replicates at 37 º.

Whether it is one species different temperatures or different species one temperature, the basic assumption for applying ANOVA is that the samples of each level comes from a 'population' having that characteristic. The sample of bacteria cultured in 25º, 30, 35, 40º, and 45º are assumed to have come from the respective populations of bacteria cultured in 25 º, 30 º, 35 º, 40 º, and 45 º. Similarly, samples of k number of species of bacteria are supposed to have come from their respective population.

The null-hypothesis to be tested in the one way classification, completely random design with fixed effects is $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$, represented by $X_1, X_2, X_3, \ldots\ldots X_k$ from $n_1, n_2, n_3 \ldots . n_k$ samples.

It is also possible to have one way classification, completely random design with random effects. In the experiments described earlier, we selected the levels of treatment, the temperatures or the number of species and therefore, the design was described to have fixed effects. Instead, if the levels of treatment are selected at random, the experimental design is said to have random effects. For example, if the

five temperatures are selected randomly from a wide range of all possible temperatures in which the bacterial can grow, then the bacteria cultured at each of the random temperatures would represent a population of bacterial growing in the temperature. Here, these populations from which the samples were drawn have themselves been selected randomly selected randomly from large number populations. Hence, the design is said to have random effect.

11.3.2 Two-way classification Design or Factorial Design

When two factors are involved in experiments, the design is referred to as two-way classification, completely random design with fixed effects or factorial experiments. In a factorial experiments design effects of two or more factors operating on a third factor is investigated. The first two factors may be in two or more levels.

For example, we may be interested in knowing the effects of temperature and pH on bacterial growth. The levels of the factor temperature may be two, say, 30º and 40 º, and the levels of the pH may also be two, say, pH4 and pH10. Application of ANOVA would enable us to separate and evaluate not only the effects of the two factors, temperature and pH, but also the interaction of these two factors. The procedure for handling the data in a factorial design of experiment is given below,

11.4 Factorial design of experiment

Let A and B are two factors affecting a specific parameter in the sample units, A1 and A2 be the levels of A and B1 and B2 be the levels of B, at which the effects is studied nA_1B_1, nA_2B_1, nA_2B_2 be the sample in the k groups and N be the total of all the samples;

Step 1: The data organized as follows

NOTES

Factor B	Factor A		sum
	Level A1	Level A2	
Level B1	$\sum_{A1B1}^{n\,A1B1} X$	$\sum_{A2B1}^{n\,A2B1} X$	$\sum_{B1} X$
Level B2	$\sum_{A1B2}^{n\,A1B2} X$	$\sum_{A2B2}^{n\,A2B2} X$	$\sum_{B2} X$
Sum	$\sum_{A1} X$	$\sum_{A2} X$	

Step 2; Add all the N values of X to get $\sum X$, square each X and Compute $\sum X^2$ (sum of the squares of all the N values)

Step 3; $SS_{(total)} = \sum X^2 - \frac{(\sum X^2)}{N}$

$$SS_A = \frac{(\sum XA1)^2}{n\,A1} + \frac{(\sum XA2)}{n\,A2} + \frac{(\sum X)^2}{N}$$

$$SS_B = \frac{(\sum XB1)^2}{n\,B1} + \frac{(\sum XB2)}{n\,B2} + \frac{(\sum X)^2}{N}$$

Step 4;

$$SS_{(interaction\ total)} = \frac{(\sum XA1B1)}{n\,A1B1} + \frac{(\sum XA2B1)}{n\,A1B2} + \frac{(\sum XA2B2)}{n\,A2B2} - \frac{(\sum X)^2}{N}$$

To remove the separate effects of A and B, SS of these factors is subtracted from SS interaction total.

NOTES

$SS_{interaction} = SS_{interaction\ total} - (SS_A + SS_B)$

Step5; $SS_{Error} = SS_{total} - SS_A + SS_B + SS_{interaction})$

Step 6; the ANOVA table is constructed as follows.

Source of variation	**DF**	**SS**	**MS**	**F**
Factor A	(2-1)	SS_A	SS_A/DF= MS_A	$F\ A = \frac{MS\ A}{MS\ e}$
Factor A	(2-1)	SS_B	SS_B/DF= MS_B	$F\ B = \frac{MS\ B}{MS\ e}$
Interaction of A and B	(2-1)	$SS_{interaction}$	SS_i/DF= MS_i	$F\ i = \frac{MS\ i}{MS\ e}$
Error (within)	(N-k)	SS_{error}	SS_e/DF= MS_e	
Total	(N-1)			

Step7; Referring to the table, critical F's required at the appropriate DF, to reject the following H0, is obtained.

1. H_{01}: $\mu A_1 = \mu A_2$
2. H_{02}: $\mu B_1 = \mu B_2$
3. H_{03}: $\mu_A = \mu_B$

If the calculated F is greater than the table F, H0 is rejected and the inference is discussed.

11.5 F-test:

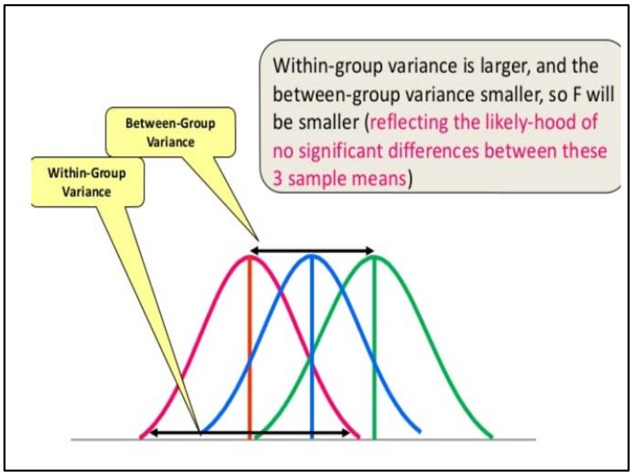
NOTES

The test measures if the means of different samples are significantly different or not is called the F-Ratio. Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.

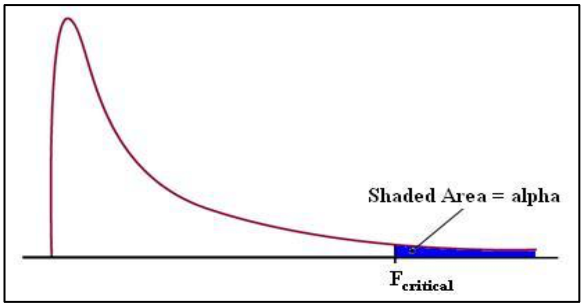
F = between group variability / within group variability.

This above formula is pretty intuitive. The numerator term in the F-statistic calculation defines the between-group variability. As we read earlier, as between group variability increases, sample means grow further apart from each other. In other words, the samples are more probable to be belonging to totally different populations.

This F-test calculated here is compared with the F-critical value for making a conclusion. In terms of our medication example, if the value of the calculated F-statistic is more than the F-critical value (for a specific α/significance level), then we reject the null hypothesis and can say that the treatment had a significant effect.



Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.



Therefore, there is only one critical region, in the right tail (shown as the blue shaded region above). If the F-statistic lands in the critical region, we can conclude that the means are significantly different and we reject the null hypothesis. Again, we have to find the critical value to determine the cut-off for the critical region. We'll use the F-table for this purpose.

We need to look at different F-values for each alpha/significance level because the F-critical value is a function of two things: df_{within} and $df_{between}$.

11.6 Steps involved in ANOVA:

We would look at the steps to follow to perform Analysis of Variance (ANOVA). This would be very clear and easy to follow. ANOVA is used to analyze the difference in the means of different groups (for 3 or more groups).

11.6.1 The Seven Steps are

NOTES

Step 1: Calculate the Mean

Step 2: Setup the null and alternate hypothesis

Step 3: Calculate the Sum of Squares

Step 4: Calculate the Degrees of Freedom

Step 5: Calculate the Mean Squares

Step 6: Calculate the F Statistic

Step 7: Look up statistical Table and state our conclusion

Step 1: Calculate all the mean

We need to calculate all the means for all the groups in the question. Then also need to calculate to overall means with all the data combined as one single group.

Step 2: Set up the null and alternate hypothesis and the Alpha

The null hypothesis assumes that there is no variance data in different groups. In other words, the means are the same.

The alternate hypothesis states the means are different. So we can state as follows:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: $\mu_1 \neq \mu_2 \neq \mu_3$

Step 3: Calculate the Sum of Squares

Calculate the Sum of Squares Total (SST): The SS $_{total}$ is the Sum of Squares (SS) based on the entire set in all the groups. In this case, treat all the data from all the groups like on single combined set of data,

$$SS_{total} = \sum_{k=0}^{n}(\overline{XJ} - \bar{X})^2$$

Calculate the Sum of Squares within Groups (SSW): the sum of squares within each group. After calculating the sum of squares for each group, then you add them together for all the groups. That is why you have the sum symbol twice in the formula,

- The internal sum is for within the group
- The external sum is to sum the sums! $\bar{X}_j$

$$SS_{within} = \sum_{j=1}^{k} \quad \sum_{j=1}^{l}(X - \bar{X}\text{j})$$

But for learning purposes, we would calculate the third one. So let’s keep moving!

Calculate the Sum of Squares between Groups (SSB): This is the sum of squares with the groups taken as single elements.

Assuming there are three groups will have to do the following:

(group1_mean – total_mean)2 + (group2_mean – total_mean)2 + group3_mean – total_mean)2

$$SS_{between} = \sum_{j=1}^{k} (\overline{XJ} - \bar{X})^2$$

Verify that

$SS_{total} = SS_{between} + SS_{within}$.

Step 4: Calculate the Degrees of Freedom (df)

NOTES

Calculate the Degrees of Freedom Total (DFT)

$$df_{total} = n - 1$$

Where n is to total of all the data sets combined

Calculate the Degrees of Freedom within Groups (DFW)

$$df_{within} = k - 1$$

K is the number of groups.

So, if there are three groups of measurements, then k = 3

Calculate the Degrees between Groups (DFB)

$$df_{between} = n - k$$

We can verify that:

$$dft = dfw + dfb$$

Step 5: Calculate the Mean Squares

Calculate the Mean Squares between (MSB)

$$MS_{between} = \frac{SSbetween}{dfbetween}$$

Calculate the Mean Squares within (MSW)

$$MS_{within} = \frac{SSwithin}{dfwithin}$$

Step 6: Create a Summary Table and Calculate the F Statistic

NOTES

This is the same table created in Step 2. If just have to fill it with actual results based on our calculations. Stating with this table makes the problem easier to solve.

Calculate the F Statistic using the formula

$$F = \frac{MSbetween}{MSwithin}$$

Step 7: Look up F from table

Look up the tabulated value of F (critical value) of F from the statistical table and compare it with the value calculated (absolute value).

If the absolute value is greater than the critical value, we reject the null hypothesis and conclude that there is significant different between the means of the populations. Otherwise, accept the null hypothesis or fail to reject the null hypothesis.

11.7 Let us sum up

In this unit you have learn ANOVA, F-test and their methods. This knowledge would make understand how we find the significance from the observed variables. Student can learn the measure of the variability of the scores in the data set.

11.8 Unit-End Exercise

1. Explain the methods of ANOVA?
2. List out the steps involved in ANOVA?
3. Define F-test.

NOTES

11.9 Check your progress:

1. What is the difference between one way ANOVA and two ways ANOVA?
2. What does a high F value mean?
3. How do you calculate the sum of squares?

11.10 Answer to check your progress:

i. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA. ... In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.

ii. A high F value means that your data does not well support your null hypothesis. ... So a large F value indicates that a linear model is more compatible with the data than a constant average model.

iii. To calculate this, subtract the number of groups from the overall number of individuals. SS_{within} is the sum of squares within groups. The formula is: degrees of freedom for each individual group (n-1) * squared standard deviation for each group.

11.11 SUGGESTED READINGS

1. Pagano, M.& Gauvreau, K (2007), Principles of Biostatistics.

2. Rao, K. V (2007), Biostatistics - A Manual of Statistical Methods for use in Health Nutrition and Anthropology.

3. Sundaram, K.R (2010), Medical Statistics-Principles & Methods, BI Publications, New Delhi

Unit XII

NOTES

12.1 Introduction

The relationship between two or more variables is called correlation, and the variables are said to be correlated. The relationship between two variables is also known as “co variation”. The term “relationship” can be used in two different senses, mutual dependence and cause and effect relationship. In this unit we discussed about correlation, types of correlation and, scatter diagram.

12.2 Objective

- Interpret the nature of relationship between two variables
- Calculate the measure of correlation
- Critically examine the degree and direction of the relationships.

12.3 Correlation

The relationship between two or more variables is called correlation. Consider the two variables, rate of oxygen consumption and metabolism in organisms. When the oxygen consumption increases, there is increase in the metabolism as well. Similarly when the organism increases its activity (metabolism) it consumes more oxygen. On the other hand, when the oxygen consumption decreases, the activity, i.e., the metabolisms decreases. When the organisms become less active, its oxygen consumption also becomes lesser. A relationship between two or more variables in which a change in the value of one the two variables brings about a change in the value of the other variable is said to be 'mutually dependent'.

12.4 Methods of study of Correlation

We can study the presence or absence and extend of correlation between variables by one of the following methods: 1) scatter diagram: 2) correlation graph: 3) Karl Pearson's coefficient of correlation (r).

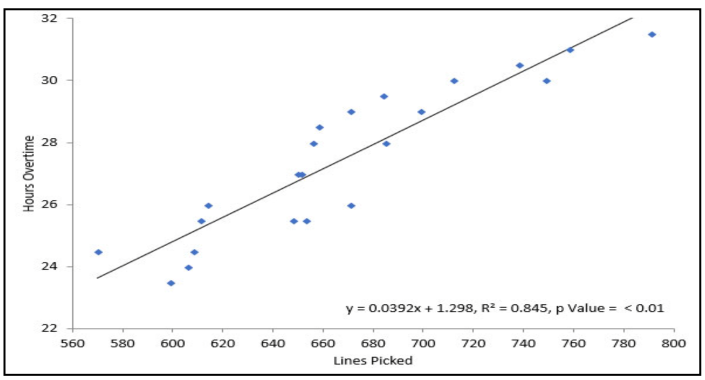
Scatter diagram and correlation graphs are graphical methods and they indicate only the nature of the correlation whether positive or negative. No numerical measure of the extent of correlation is given by these methods. The Karl pearson's coefficient of correlation gives the magnitude of correlation between two variables in numerical terms.

NOTES

12.4.1 Scatter diagram

It is an easy and simple method of studying correlation between two variables. Scatter diagram is constructed as follows. If X and Y are pairs of variables, the values of the variables X are marked in the X-axis and the values of variables Y are marked in the Y-axis. A point is plotted against each value of X and the corresponding Y value. A swarm of dots is obtained, and this is called the scatter diagram. The nature of the scatter of dots in the diagram gives an idea of the nature of correlation between the variables given, as shown in figure.

If the plotted dots from a straight line running left to right in the upward direction the correlation is perfectly positive.



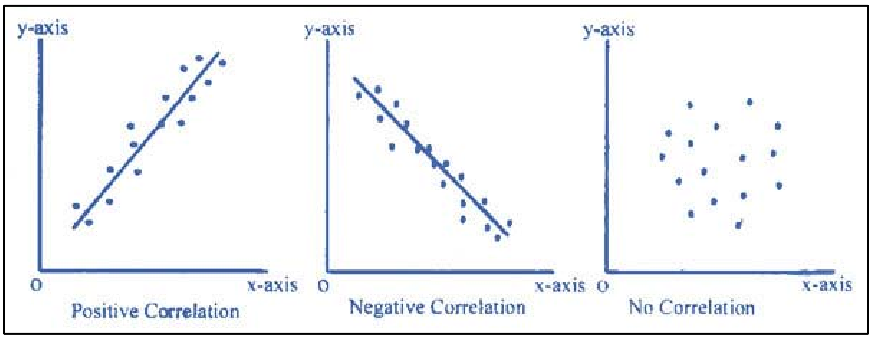
If the dots are scattered around a straight line running from left to right in an upward direction, then the correlation between the two variables is positive.

NOTES

If the dots of the scatter diagram from a straight line running from left to right in the downward direction, the correlation between the two variables is perfect negative.

A scatter diagram in which the plotted dots from a swarm around a straight line that runs from left to right in the downward direction indicates negative correlation between the variables.

In cases where there is no correlation between the variables, the scatter of dots in the diagram will not from either a straight line or even a flow of dots from left to right in the upward or downward direction.

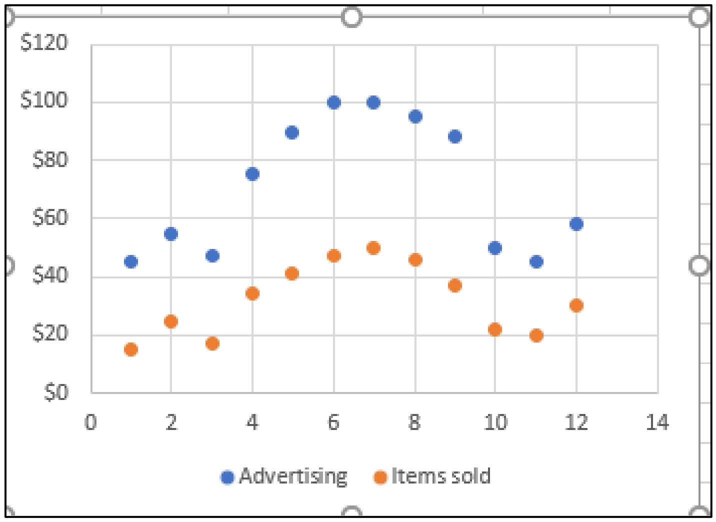


12.4.2 Correlation Graph

When the given variables are with reference to a period of time, correlation graph is the ideal method to understand the relationship between the two variables. To draw the correlation graph, the period of time is marked on the X-axis and the values of the two variables on the Y-axis or if necessary on both the Y-axis. The points of a variable plotted against time are joined by a line to form

NOTES

a curve. Different curves are constructed for each of the two variables.



Correlation graph showing positive correlation between two variables

In a correlation graph, if the curves of the two variables are close to each other and if they move in the same direction, the variables have a positive correlation. On the other hand, if the curves of the two variables move in opposite directions, the variables are negatively correlated.

12.4.3 Karl Pearson's Coefficient of Correlation

Coefficient of Correlation(r) is a quantitative measure of the correlation between two variables. The correlation coefficient is calculated on the basis of the following assumptions about the two variables. 1) The correlation between the two variables is linear. 20 the two variables are related to each other in a cause and effect

relationship. 3) The values of the two variables are affected by factor that is common to both the variables. The equation for getting Karl pearson's coefficient(r) is

$$r = \frac{\sum(X - X)(Y - Y)}{nS_{\chi}S_{\gamma}}$$

Where,

R= Coefficient of correlation

X= variable X

X= Mean of variable X

Y= variable Y

Y= Mean of variable y

n = number of paris of variables

S_{χ}= SD of variables X and

S_{γ}= SD of variables of Y

A modification of the above basic formula that is easier to use is as follows

$$r = \frac{\sum XY - \sum X \sum Y/n}{\sqrt{[\sum X^2 - (\sum X)^2][\sum Y^2 - (\sum Y)^2/n}}$$

Any scientific calculator will give the value of r if you feed the bivariate data, i.e., values of the variables X and Y. it will also give

NOTES

all the necessary statistics necessary to apply in the second formula.

12.4.3.1 Interpretation of the Karl Pearson's Coefficient of Correlation

The value of r always lies between +1 and -1,

When;

r = +1, the correlation between the two variables is period positive;
r = -1, the correlation is perfect negative;
r = a positive value lying between 0 and +1, the correlation is positive:
r = a negative value lying between -1 and 0, the correlation is negative:
and,
r = 0, there is no linear correlation between the two given variables.

An important thing to be understood in the coefficient of correlation (r) is that it does not represent a percent agreement between the two given variables. However if r = -1 and +1 or 0, there might be 100 percent agreement. But if r = 0.5 or any other value, it does not indicate 50% or any other respective percent agreement between the two variables. The percent agreement between two variables increases exponentially with the increasing r values.

12.5 Rank Correlation

NOTES

Sometimes there doesn't marked a linear relationship between two random variables but a monotonic relation (if one increases, the other also increases or instead, decreases) is clearly noticed. Pearson's Correlation Coefficient evaluation, in this case, would give us the strength and direction of the linear association only between the variables of interest. Herein comes the advantage of the Spearman Rank Correlation methods, which will instead, give us the strength and direction of the monotonic relation between the connected variables. This can be a good starting point for further evaluation.

12.5.1 The Spearman Rank-Order Correlation Coefficient

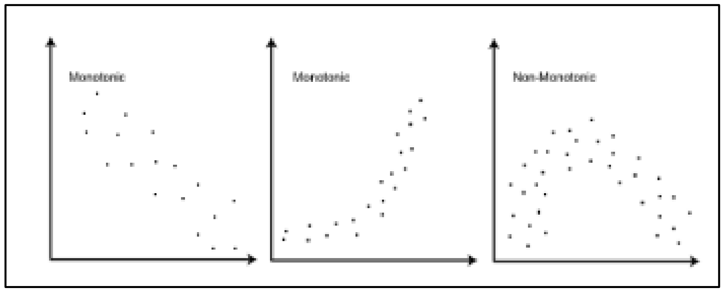
The Spearman's Correlation Coefficient, represented by ρ or by rR, is a nonparametric measure of the strength and direction of the association that exists between two ranked variables. It determines the degree to which a relationship is monotonic, i.e., whether there is a monotonic component of the association between two continuous or ordered variables.

Monotonicity is "less restrictive" than that of a linear relationship. Although monotonicity is not actually a requirement of Spearman's correlation, it will not be meaningful to pursue Spearman's correlation to determine the strength and direction of a monotonic relationship if we already know the relationship between the two variables is not monotonic.

On the other hand if, for example, the relationship appears linear (accessed via scatterplot) one would run a Pearson's correlation

NOTES

because this will measure the strength and direction of any linear relationship. Monotonicity



12.5.2 Spearman Ranking of the Data

We must rank the data under consideration before proceeding with the Spearman's Rank Correlation evaluation. This is necessary because we need to compare whether on increasing one variable, the other follows a monotonic relation (increases or decreases regularly) with respect to it or not.

Thus, at every level, we need to compare the values of the two variables. The method of ranking assigns such 'levels' to each value in the dataset so that we can easily compare it.

Assign number 1 to n (the number of data points) corresponding to the variable values in the order highest to lowest.

In the case of two or more values being identical, assign to them the arithmetic mean of the ranks that they would have otherwise occupied.

NOTES

For example, Selling Price values given: 28.2, 32.8, 19.4, 22.5, 20.0, 22.5 The corresponding ranks are: 2, 1, 5, 3.5, 4, 3.5 The highest value 32.8 is given rank 1, 28.2 is given rank 2,…. Two values are identical (22.5) and in this case, the arithmetic means of ranks that they would have otherwise occupied ($\frac{3+4}{2}$) has to be taken.

12.5.2.1 The Formula for Spearman Rank Correlation

$$r_R = 1- \frac{6\sum idi^2}{n(n^2-1)}$$

Where n is the number of data points of the two variables and is the difference in the ranks of the ith element of each random variable considered. The Spearman correlation coefficient, ρ, can take values from +1 to -1.

- A ρ of +1 indicates a perfect association of ranks
- A ρ of zero indicates no association between ranks and
- ρ of -1 indicates a perfect negative association of ranks.
- The closer ρ is to zero, the weaker the association between the ranks.

12.5.2.2 Solved Examples for On Spearman Rank Correlation

The following table provides data about the percentage of students who have free university meals and their CGPA scores. Calculate the Spearman's Rank Correlation between the two and interpret the result.

NOTES

State University	% of students having free meals	% of students scoring above 8.5 CGPA
Pune	14.4	54
Chennai	7.2	64
Delhi	27.5	44
Kanpur	33.8	32
Ahmedabad	38.0	37
Indore	14.9	68

Solution: Let us first assign the random variables to the required data –

X – % of students having free meals

Y – % of students scoring above 8.5 CGPA

Before proceeding with the calculation, we'll need to assign ranks to the data corresponding to each state university. We construct the table for the rank as below,

NOTES

State University	**d_X = $Ranks_X$**	**d_Y = $Ranks_Y$**	**d = (d_X – d_Y)**	**d^2**
Pune	3	4	-1	1
Chennai	2	6	-4	16
Delhi	5	3	2	4
Kanpur	6	1	5	25
Ahmedabad	7	2	5	25
Indore	4	7	-3	9
Guwahati	1	5	-4	16
				$\Sigma d^2 = 96$

Now, using the formula (with n = 7 here) –

$$rR=1–\frac{6\Sigma idi2}{n(n2-1)}$$

$$= 1–\frac{6.96}{7.(49-1)}$$

NOTES

$$= 1 - \frac{576}{336}$$

$$\mathbf{= -0.714}$$

Such a strong negative coefficient of correlation gives away an important implication.

12.6 Types of correlation

Correlation between variables may be simple or multiple. A simple correlation deals with only two variables where as a multiple correlation deals with more two variables.

Correlation between two variables may be a positive correlation or a negative correlation. Whether it is positive or negative, it may be linear or non-linear.

12.6.1 Positive correlation

A Correlation between two variables in which, with an increase in the values of one variable the values of the other variable also increases, and with a decrease in the value of the one variable the value of the other variable also decreases, is said to be a positive correlation. In other words, in a positive correlation between two variables the values of both the variables move in the same direction. For example, the correlation between the environmental temperature and the body temperature of poikolotherms is a positive correlation.

12.6.2 Negative correlation

A correlation between two variables in which when there is an increase in the values of one variable, the values of the other variable decreases, and when there is decrease in the values of one variable the other variable increases, is said to be a negative correlation. In other words, in a negative correlation the values of the two variables move in opposite direction. For example, the correlation between environmental temperature and bacterial growth, having a cause and effective relationship, is negative one. With an increase in the temperature the bacterial growth declines and with a decrease in the temperature the bacterial growth increases.

12.6.3 Linear correlation

When the values of two variables vary in a constant ratio, the correlation between the two variables is said to be linear. The correlation between the optical density and the intensity of the colour of a solution is an example of linear correlation.

12.6.4 Non-linear or Curvilinear Correlation

If the amount of change in the value of one variable and the corresponding amount of change in the other variable are not in constant ratio, the correlation between the two variables is said to be non-linear or Curvilinear. The correlation between the length and weight of fish is generally a non-linear correlation.

NOTES

12.7 Let us sum up

In this unit study a many effective ways of forecasting and predicting possible outcomes based on past observations, through other statistical methods too need to be implemented to get a complete picture of a the situation.

12.8 Unit-End Exercise

1. What is correlation and explain the methods to study correlation?
2. Define scatter diagram with the graphical representation?
3. What is Spearman Ranking of the Data with its formulae?
4. List out the types of correlation?

12.9 Check your process

1. What are the different methods of correlation?
2. What is Pearson r formula?
3. What are 3 types of correlation?

12.10 Answers to check your process

1. Methods of Determining Correlation:

- Scatter Diagram Method.
- Karl Pearson's Coefficient of Correlation.
- Spearman's Rank Correlation Coefficient; and.

2. Pearson r correlation formula: The following formula is used to calculate the Pearson r correlation: rxy = Pearson r correlation coefficient between x and y. n = number of observations. xi = value of x (for ith observation)

NOTES

3. There are three types of correlation: positive, negative, and none (no correlation). Positive Correlation: as one variable increases so does the other. Height and shoe size are an example; as one's height increases so does the shoe size. Negative Correlation: as one variable increases, the other decreases.

12.11 SUGGESTED READINGS

1. Draper, N.R. and Smith, H (2003). Applied Regression Analysis, John Wiley & Sons. Rossi R.J.(2010).Applied Biostatistics for Health Sciences, Wiley.

2. Gurumani,N., (2015).“An Introduction to Biostatistics”, MJP Publisher, 2nd Edition

3. Warren, J., Gregory, E. and Grant, R. (2004) “Statistical methods in Bioinformatics”; First edition, Springer-Verlag, Berlin.

NOTES

UNIT XIII

NOTES

13.1 Introduction

If two variables, X and Y, are related to one another with a significant correlation, it is possible to obtain an estimate of the value X and Y is known, and the value of Y, if X is known. The statistical device with the help of which the unknown values of a variable from the known values of another correlated variable is estimated is called regression. A line that gives the best estimate of the value of one variable when the value of the other variable is given is known as regression line.

NOTES

Two regression lines X on Y and Y on X, can be drawn for a series of bivariate data. The regression line of X on Y gives the best estimate the value of X when a value of Y is given. The regression line of Y on X gives the best estimate of the value of Y when the value of X is given.

Regression lines are given algebraic expression, the regression equations. Regression equations are used to draw the regression lines. They are also used as numerical methods to find out the best estimate of values of the variables from the values of the other variables.

13.2 objectives

- Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
- •Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
- The applications areas are in 'explaining' variations in sales of a product based on advertising expenses, or number of sales people, or number of sales offices, or on all the above variables.

13.3 Regression Equation

For every series of bivariate data two regression equations are derived, viz., regression equation of X on Y that is used to draw

NOTES

the regression line of X on Y, and the regression equation of Y on X that is used to draw the regression line of Y on X. The regression equation of X on Y is as follows,

$$\frac{(X - \bar{X}) = rS_{\chi}\,(Y - \bar{Y})}{S_{Y}}$$

The regression equation of Y on X is as follows.

$$\frac{(Y - \bar{Y}) = rS_{\chi}\,(X - \bar{X})}{S_{Y}}$$

In the above equations, $\bar{X}$ is the mean of variable X, $\bar{Y}$ is mean of variable Y, S_{χ} and S_{γ} are the SD's of the variables X and Y, respectively and r = is the correlation coefficient between the variables X and Y.

If a value of Y is given, the corresponding value of X can be estimated using the equation (1). If a value of X is given, the corresponding value of Y can be estimated using equation (2). Equation (1) yields a final form as X = AY + B, and the equation (2), Y = AX + B.

In the expression X = AY + B, X is considered the dependent variable and Y is an independent variable. X is said to be a function of Y. "A" in the above expression is called the regression coefficient or slope of the regression equation of X on Y. "B" in the expression is referred to as the constant or intercept.

Similarly, in the expression Y = AX + B, Y is the dependent variable and X is the independent variable, and Y is said to be a

NOTES

function of X. the “A” is the regression coefficient or slope and “B”, the constant of the regression equation of Y on X.

Suppose there is a unit change in the value of \Y, the corresponding change in the value of X is given by the regression coefficient of X on Y> that is, when the is unit change in the value of Y, the value of X will change by the amount rS_χ/S_γ In other words, regression coefficient decides what should be the slope of the regression line of X on Y.

Similarly, the regression coefficient of Y on X, r S_γ/S_χ, gives the amount of change in the value of Y when there is a unit change in the value of X and, thus describing the slope of the regression line of Y on Y.

The formula for obtaining the regression coefficient and constant of the two regression equations are given below,

13.4 Regression Coefficient Constant Intercept of Regression Equation

Regression equation of X on Y,

$(X - \bar{X}) = \frac{S_X}{S_Y}(Y - \bar{Y})$

X = AY + B

A = coefficient of Y in the expression equation of on Y and therefore the regression coefficient of X.

$A = \frac{S_X}{S_Y}$ or $A = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - (\sum Y)^2}$

NOTES

B = constant

$$B = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{n\sum Y^2 - (\sum Y)^2}$$

Regression equation of Y on X

A = coefficient of X in the regression equation of Y on X and therefore the regression coefficient of Y.

$$A = \frac{rS_Y}{rS_X}$$

$$A = \frac{n \sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$B = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

The derivation of the two regression equations,

Example, Obtain the two regression equations, length (X) on weight (Y) and weight (Y) on length (X) from the following data on the length (X in cm) and weight (Y in g) of fish.

X	**5**	**7**	**3**	**1**	**9**	**12**	**8**	**3**
Y	**8**	**9**	**5**	**4**	**9**	**13**	**7**	**9**

Step1: calculation of $\bar{X}$, SD of X, S_χ, -$\bar{Y}$, SD of Y, S_γ and the correlation coefficient between X and Y. these statistics have already been calculated in the below,

X length in cm	Y weight in g

NOTE

n = 8	
X = 6cm	Y = 8g
S_χ = 3.42 cm	S_Y = 2.6 g
r = 0.8	

Step2: Regression equation of X on Y, i.e., length on weight.

$$(X-\bar{X}) = \frac{rS_\chi}{S_Y}(Y-\bar{Y})$$

$$(X-6) = \frac{0.8(3.42)}{2.6}(Y-8)$$

$$(X-6) = \frac{2.736}{2.6}(Y-8)$$

(X-6) = 1.05 (Y-8)

(X-6) = 1.05Y -8(1.05)

(X-6) = 1.05Y-8.40

X = 1.05Y- 8.4 + 6

X = 1.05Y – 2.40

In the regression equation of X = 1.05 Y-2.40, 1.05 is the regression coefficient or the slope, and 2.40 is the constant.

Step3: Regression equation of Y on X, i.e., weight on length.

$$(Y-Y) = \frac{rS_\chi}{S_Y}(X-X)$$

$$(Y-8) = \frac{0.8(2.6)}{3.42}(X-6)$$

$$(Y-8) = \frac{2.08}{3.42}(X-6)$$

NOTES

(Y-8) = 0.61(X-6)

(Y-8) = 0.61X – (0.61) (6)

(Y-8) = 0.61X -3.66

X = 0.61X - 3.66 + 8

X = 0.61X – 4.34

In the above regression equation of Y on X, 0.61 is the regression coefficient or slope, and 4.34 is the constant. Using the above two regression equations we can estimate the values of X from the known values of Y and Y from the known values of X, and construct the two regression lines, as shown.

13.5 Types of regression analysis:

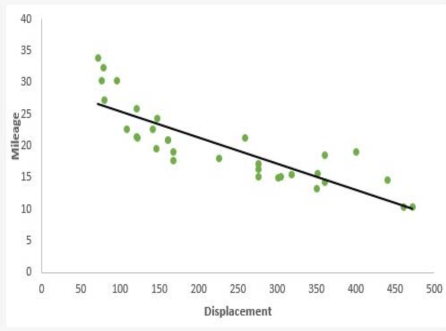
Every regression technique has some assumptions attached to it which we need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution.

13.5.1. Linear Regression

It is the simplest form of regression. It is a technique in which the dependent variable is continuous in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.We can observe that the given plot represents a somehow linear relationship between the mileage and

NOTES

displacement of cars. The green points are the actual observations while the black line fitted is the line of regression.



Regression Analysis

Only 1 independent variable and 1 dependent variable, it is called simple linear regression.

More than 1 independent variable and 1 dependent variable, it is called multiple linear regression.

13.5.1.1 Assumptions of linear regression:

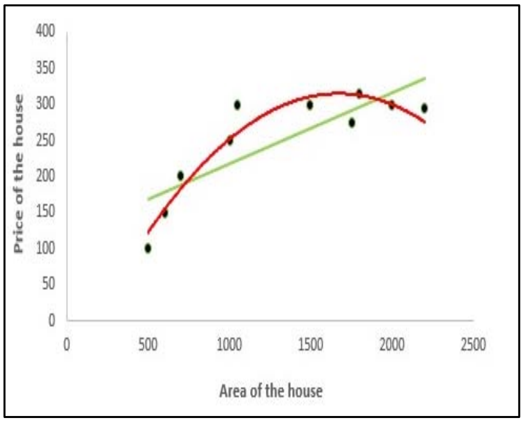
1. There must be a linear relation between independent and dependent variables.
2. There should not be any outliers present.
3. No heteroscedasticity.
4. Sample observations should be independent.
5. Error terms should be normally distributed with mean 0 and constant variance.

13.5.2. Polynomial Regression

NOTES

It is a technique to fit a nonlinear equation by taking polynomial functions of independent variable.

In the figure given below, you can see the red curve fits the data better than the green curve. Hence in the situations where the relation between the dependent and independent variable seems to be non-linear we can deploy Polynomial Regression Models.



13.5.3. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of

NOTES

the most commonly used Stepwise regression methods are listed below:

- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the methods to handle higher dimensionality of data set.

13.5.4. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represent by following equation.

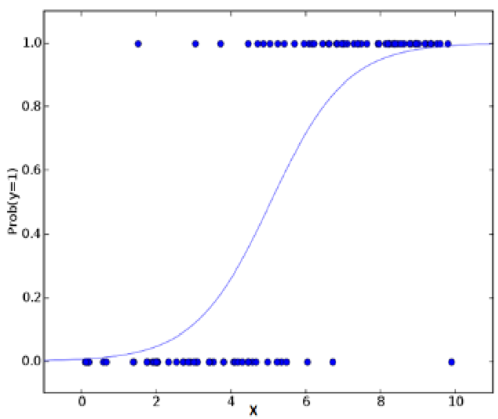
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is logit function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).



13.5.4.1 Important Points:

It is widely used for classification problems

Logistic regression doesn’t require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio

NOTES

To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression

It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square.

13.5.5. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

13.5.5.1 Important Points:

The assumptions of this regression are same as least squared regression except normality is not to be assumed.

It shrinks coefficients to zero (exactly zero), which certainly helps in feature selection.

This is a regularization method and uses.

If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero.

NOTES

13.6 Graphical Method

It involves drawing a scatter diagram with independent variable on X-axis and dependent variable on Y-axis. After that a line is drawn in such a manner that it passes through most of the distribution, with remaining points distributed almost evenly on either side of the line.

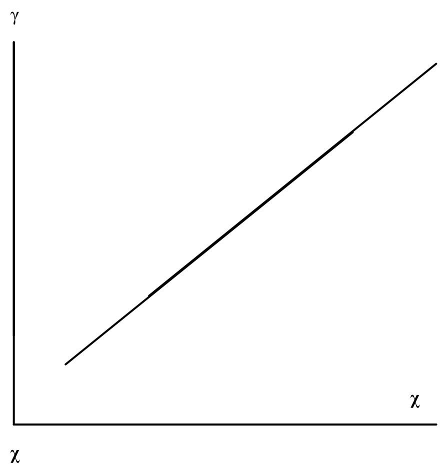
A regression line is known as the line of best fit that summarizes the general movement of data. It shows the best mean values of one variable corresponding to mean values of the other. The regression line is based on the criteria that it is a straight line that minimizes the sum of squared deviations between the predicted and observed values of the dependent variable.

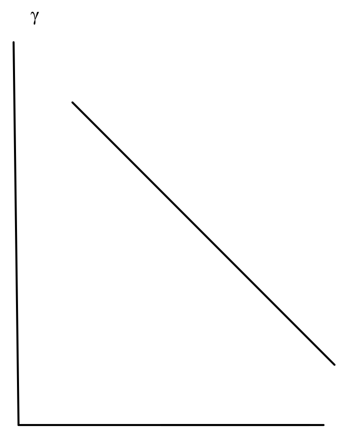
one or two regression lines are drawn on a graph paper to estimate the values of one variable say, X on the basis of the given values of another variable say, Y.

If there is a perfect correlation (i.e. r =1) between the two variables, only one regression line can be drawn, for in that case both the regression lines of X on Y, and Y on X will coincide. In case, the correlation between the two variables is not perfect, two lines of regression are to be drawn on the graph paper one of which will be the regression line of X on Y, and the other will be the

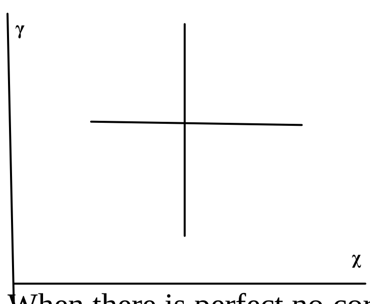
NOTES

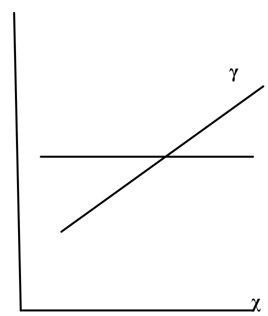
regression line of Y on X. If there is no correlation between the two variables, then the two line of regression will be perpendicular to each other. The different forms of the regression line undr different state of correlation are exhibited hare as under:





When there is perfect positive correlation. i.e. r = 1 When there is perfect negative correlation. 1 i.e. r = -1





When there is perfect no correlation. i.e. r = 0 When there is high degree of correlation.

13.7 Algebraic Method

NOTES

The algebraic method refers to various methods of solving a pair of linear equations, including graphing, substitution and elimination.

The graphing method involves graphing the two equations. The intersection of the two lines will be an x,y coordinate, which is the solution.

With the substitution method, rearrange the equations to express the value of variables, x or y, in terms of another variable. Then substitute that expression for the value of that variable in the other equation.

For example, to solve:

$8x+6y=16$

$-8x-4y=-8$

First, use the second equation to express x in terms of y:

{-8} x=-8+4yx= $\frac{-8+4y}{-8x}$1-0.5y

Then substitute 1 - 0.5y for x in the first equation:

$8(1-0.5y)+6y=16$

$8-4y+6y=16$

$8+2y=16$

$2y=8$

NOTES

$y=4$

Then replace y in the second equation with 4 to solve for x:

$8x+6(4) =16$

$8x+24=16$

$8x=-8$

$x=-1$

The second method is the elimination method. It is used when one of the variables can be eliminated by either adding or subtracting the two equations. In the case of these two equations, we can add them together to eliminate x:

$8x+6y=16$

$-8x-4y=-8$

$0+2y=8$

$y=4$

Now, to solve for x, substitute the value for y in either equation:

$8x+6y=16$

$8x+6(4)=16$

$8x+24=16$

$8x+24-24=16-24$

$8x=-8$

$x = -1$

NOTES

13.8 Regression Analysis Is Important

The importance of regression analysis is that it is all about data: data means numbers and figures that actually define your business. The advantages of regression analysis is that it can allow you to essentially crunch the numbers to help you make better decisions for your business currently and into the future.

The regression method of forecasting means studying the relationships between data points,

- Predict sales in the near and long term.
- Understand inventory levels.
- Understand supply and demand.
- Review and understand how different variables impact all of these things.

Companies might use regression analysis to understand, for example:

- Why customer service calls dropped in the past year or even the past month.
- Predict what sales will look like in the next six month.
- Whether to choose one marketing promotion over another.
- Whether to expand the business or create and market a new product.

13.8.1 Importance of regression analysis:

NOTES

The Five Applications of Regression Analysis,

The regression analysis method of forecasting generally involves five basic applications. There are more, but businesses that believe in the advantages of regression analysis generally use the following:

138.1.1 Predictive analytics: This application, which involves forecasting future opportunities and risks, is the most widely used application of regression analysis in business. For example, predictive analytics might involve demand analysis, which seeks to predict the number of items that consumers will purchase in the future. Using statistical formulas, predictive analytics might predict the number of shoppers who will pass in front of a given billboard and use then use that information to place billboards where they will be the most visible to potential shoppers. And, insurance companies use predictive analysis to estimate the credit standing of policyholders and a possible number of claims in a given time period.

13.8.1.2 Operation efficiency: Companies use this application to optimize the business process. For example, a factory manager might use regression analysis to see what the impact of oven temperature will be on loaves of bread baked in those ovens, such as how long their shelf life might be. Or, a call center can use regression analysis to see the relationships between wait times of callers and the number of complaints they register. This kind of data-driven decision-making can eliminate guesswork and make

the process of creating optimum efficiency less about gut instinct and more about using well-crafted predictions based on real data.

13.8.1.3 Supporting decisions: Many companies and their top managers today are using regression analysis (and other kinds of data analytics) to make an informed business decision and eliminate guesswork and gut intuition. Regression helps businesses adopt a scientific angle in their management strategies. There is actually, often, too much data literally bombarding both small and large businesses. Regression analysis helps managers sift through the data and pick the right variables to make the most informed decisions

13.8.1.4 Correcting errors: Even the most informed and careful managers do make mistakes in judgment. Regression analysis helps managers, and businesses in general, recognize and correct errors. Suppose, for example, a retail store manager feels that extending shopping hours will increase sales. Regression analysis may show that the modest rise in sales might not be enough to offset the increased cost for labor and operating expenses (such as using more electricity, for example). Using regression analysis could help a manager determine that an increase in hours would not lead to an increase in profits. This could help the manager avoid making a costly mistake

13.8.1.5 New Insights: Looking at the data can provide new and fresh insights. Many businesses gather lots of data about their customers. But that data is meaningless without proper regression

NOTES

analysis, which can help find the relationship between different variables to uncover patterns. For example, looking at the data through regression analysis might indicate a spike in sales during certain days of the week and a drop in sales on others. Managers could then make adjustments to compensate, such as making sure to maintain stock on those days, bringing in extra help, or even ensuring that the best sales or service people are working on those days.

13.9 Let us sum up

In this unit primarily it is based on quantifying the changes in the dependent variable (regressed variable) due to the changes in the independent variable by using the data on the dependent variables. This is because all the regression models whether linear or non-linear, simple or multiples relate the dependent variable with the independent variables.

13.10 Unit-End Exercise

1. What is Regression Equation?
2. Explain the types of regression analysis?
3. Define Lasso Regression?
4. Suggest your valuable comments that why Regression Analysis Is Important?

NOTES

13.11 Check your progress

1. What is an example of regression?
2. What are regression and its formula?
3. What are different types of regression?

13.12 Answers to check your progress:

1. For example, linear regression can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). ... This post will show you examples of linear regression, including an example of simple linear regression and an example of multiple linear regressions.

2. The Regression Equation is the algebraic expression of the regression lines.... Regression Equation of Y on X: This is used to describe the variations in the value Y from the given changes in the values of X.

3. Types of regression.

- Linear regression. ..
- Ridge regression. ...

13.13 SUGGESTED READINGS

1. Draper, N.R. and Smith, H (2003). Applied Regression Analysis, John Wiley & Sons.

2. Rossi R.J (2010).Applied Biostatistics for Health Sciences, Wiley

3. Zar, J.H. (1984) “Bio Statistical Methods; Prentice Hall International Edition, USA

NOTES

UNIT XIV

14.0 Introduction:

Software's is a set of programs designed to perform a well-defined function. Software programs can help by managing and analyzing the quantitative or qualitative data's. Statistical software sets of derived mathematical equation programs that are used for statistical analysis in collection, organization, analysis, interpretation and presentation of data. Statistical software's aided with programming also used in the analysis econometrics too. In modern world the development of technologies increases day by day for the well fare of human beings, as in same path this statistical software's also plays an important key role as same aspects. Statistical software are used in the various interpretations fields like

- Health researchers,
- Data analysis,
- Data miners processing,
- Data documentation,
- case selection,

NOTES

- create derived data,
- Perform file reshaping and
- Analyzing of survey data etc.

In this statistical software we mainly discuss about the software like SPSS, Model Analyst's Toolkit (MAT), The Charles River Analytics Solution, Wizard Mac and NCSS are discussed mainly for the welfare of students to know the basic knowledge in field of bioinformatics.

14.1 Objectives:

- Statistical analysis is used in exploring, presenting a large amount of data's in various fields for understanding the scientific outcomes.
- Statistical analysis may also help the users to uncover business opportunities, increase their business revenues and profits.
- These also help the health researcher to identify and interpret with new findings related with statistical analysis.

14.2 SPSS:

SPSS is short for Statistical Package for the Social Sciences, and it's used by various kinds of researchers for complex statistical data analysis.

The SPSS software package was created for the management and statistical analysis of social science data. It was originally launched in 1968 by SPSS Inc., and was later acquired by IBM in 2009.

Officially dubbed IBM SPSS Statistics, most users still refer to it as SPSS. As the world standard for social science data analysis, SPSS is widely coveted due its straightforward and English-like command language and impressively thorough user manual.

SPSS is used by market researchers, health researchers, survey companies, government entities, education researchers, marketing organizations, data miners, and many more for the processing and analyzing of survey data.

NOTES

While Survey Gizmo has powerful built-in reporting features, when it comes to in-depth statistical analysis researchers consider SPSS the best-in-class solution.

Most top research agencies use SPSS to analyze survey data and mine text data so that they can get the most out of their research projects.

14.2.1 The Core Functions of SPSS

SPSS offers four programs that assist researchers with their complex data analysis needs.

14.2.2 Statistics Program

SPSS's Statistics program provides a plethora of basic statistical functions, some of which include frequencies, cross tabulation, and bivariate statistics.

14.2.3 Modeler Program

SPSS's Modeler program enables researchers to build and validate predictive models using advanced statistical procedures.

Text Analytics for Surveys Program

SPSS's Text Analytics for Surveys program helps survey administrators uncover powerful insights from responses to open ended survey questions.

14.2.4 Visualization Designer

NOTES

SPSS's Visualization Designer program allows researchers to use their data to create a wide variety of visuals like density charts and radial boxplots with ease.

In addition to the four programs mentioned above, SPSS also provides solutions for data management, which allow researchers to perform case selection, create derived data, and perform file reshaping.

SPSS also offers the feature solution of data documentation, which allows researchers to store a metadata dictionary. This metadata dictionary acts as a centralized repository of information pertaining to data such as meaning, relationships to other data, origin, usage, and format.

There are a handful of statistical methods that can be leveraged in SPSS, including:

Descriptive statistics, including methodologies such as frequencies, cross tabulation, and descriptive ratio statistics.

Bivariate statistics, including methodologies such as analysis of variance (ANOVA), means, correlation, and nonparametric tests.

Prediction for identifying groups, including methodologies such as cluster analysis and factor analysis.

14.2.5 The Benefits of Using SPSS for Survey Data Analysis:

Analyzing statistical data, SPSS is an extremely powerful tool for manipulating and deciphering survey data.

Fun fact: The data from any survey collected via Survey Gizmo can be exported to SPSS for detailed analysis.

Exporting survey data to SPSS's proprietary .SAV format makes the process of pulling, manipulating, and analyzing data clean and easy.

By doing so, SPSS will automatically set up and import designated variable names, variable types, titles, and value labels, meaning that minimal legwork is required from researchers.

Once survey data is exported to SPSS, the opportunities for statistical analysis are practically endless.

In short, remember to use SPSS when you need a flexible, customizable way to get super granular on even the most complex data sets. This gives you, the researcher, more time to do what you do best and identify trends, develop predictive models, and draw informed conclusions.

14.2.6 For more information on the benefits of using SPSS to conduct survey data analysis, here are some helpful resources:

- Survey Gizmo's SPSS Export Documentation
- Creating and Manipulating Survey Variables Using SPSS
- SPSS Variable and Value Totals: A Quick Tutorial
- SPSS Analysis Tips: The Temporary Command
- SPSS Expert: Recording States into Regions Using SPSS
- Converting Survey Variables with SPSS Syntax

14.3 Model Analyst's Toolkit (MAT)

One of the most powerful things scientists can do to help analyze and validate scientific theories is to create models that effectively describe the world around us. Models help scientists organize their

NOTES

theories and suggest additional experiments to run. Validated models also help others in more practical applications. For instance, in the hands of military decision makers, human social cultural behavior (HSCB) models can help predict instability and the socio-political effects of missions, whereas models of the human brain and mind can help educators and trainers create curricula that more effectively improve the knowledge, skills, and abilities of their pupils.

14.4 The Charles River Analytics Solution

Under an effort for the Office for Naval Research (ONR), Charles River Analytics developed the Model Analyst's Toolkit (MAT) to specifically address the shortcomings of existing software tools. MAT is a software application for analyzing and validating scientific theories and models using real-world data. MAT has also been used on a variety of research programs to help analyze and extract meaning from complex data. A few examples are:

Evaluating sensor data from medical manikins and laparoscopic surgery training systems to support training doctors to use advanced medical devices

Evaluating physiological data from training systems to identify periods of trainee stress and excessive workload

Identifying relationships in social/political/economic data gathered by the World Bank

MAT enables social scientists to visualize and explore data, develop causal models, and validate those models against available data, as

NOTES

displayed in the figure below. MAT is particularly useful for analyzing and validating scientific theories and models of phenomena that happen over time, such as changes in unemployment, changes in crime rates, changes in financial markets, or changes in physiological measures like heart rate. It combines cutting-edge automated data analysis capabilities with a graphical user interface that together help scientists understand and explore their data and how their models do and do not match that data. It is natural and intuitive to use, even for users without mathematical or computational training. It can be used independently specific computational modeling frameworks. Specifically, scientists can use MAT to:

- Analyze many types of important scientific models
- Support easy data exploration
- Validate models against data for independent validation of models

NOTES



14.5 Wizard Mac

In Wizard Mac, no typing or programming is required for data analysis. Any professional can start their survey with the help of Wizard Mac. The predictive models help to make the business choices very easy.

14.5.1 Features & Benefits:

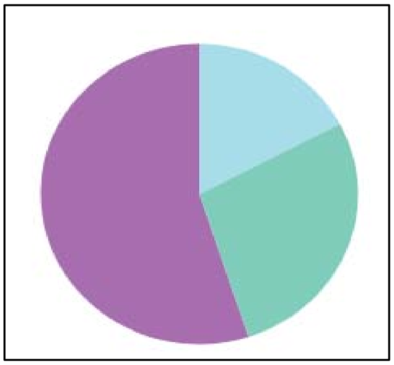
14.5.1.1 <u>Features:</u>

- Wizard Mac provides you with simple graphical results which help you understand and analyze the results easily.
- It can predict the results from the one or more selected options.
- Wizard Mac can make specific predictions after creating the predictive model with the help of sliders and pop-up buttons

14.5.1.2 Benefits

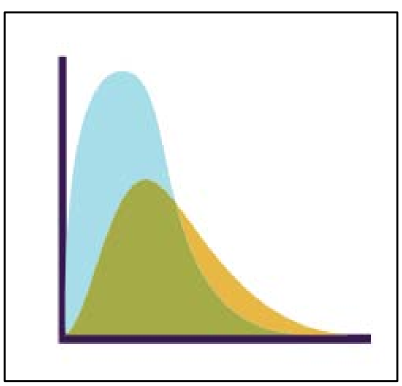
14.5.1.2.1 Elegant graphical results

Wizard is built around pictures: pictures of your data, and pictures of statistical values. The innovative graphics will help you understand data quickly and explain statistical concepts. Textbook-style illustrations — based on your data — help p-values and confidence intervals spring to life. Control-click and choose a crisp PDF or a web-ready PNG.



14.5.1.2.2 Instantaneous statistical tests

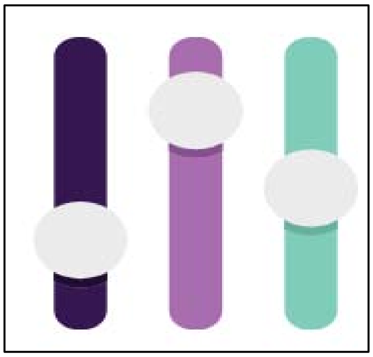
Wizard's multi-core engine reports results in the blink of an eye. Compare means with a t-test, or survival curves with a log-rank. Test for normality with a Shapiro-Wilk. It's as easy as clicking and looking at the "Bottom Line," which reports critical values and test statistics. If you don't know the test want, that's okay, too — Wizard can figure it out based on data.



NOTES

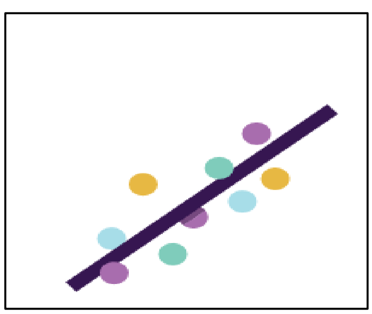
14.5.1.2.3 Advanced multivariate modeling

Wizard isn't just for quick summaries and correlations; it predicts outcomes based on one or more variables are choosing. Medical researchers can predict key probabilities using logistic, negative binomial, and proportional hazards models. Marketers can predict consumer choices with a multinomial logit or ordered probit. Or start with a simple linear model and then tinker.



14.5.1.2.4 Interpret models with ease

After creating a predictive model, use the sliders and pop-up buttons to define hypothetical scenarios and start making concrete predictions. All of your coefficients will appear together in a single equation. With one click, export your model as an interactive spreadsheet that anyone can open using Excel or Libre Office. Predictive analytics has never been easier.



NOTES

14.6 NCSS

A large number of statistical and graphical tools to analyze data are available on NCSS software. It provides facilities like organized documentation, free training videos and a 24/7 email support from the software.

14.6.1 Features:

- If can import or export data using the Data window. Quick and easy numeric results can be obtained within a few steps using NCSS.
- You can manage the data with the help of filtering and alteration features of NCSS.
- It is easy to choose the best analytical procedure for your data by using the drag and drop menu, the procedure search or the category tree.
- The personalized plot can be designed according to the requirement of the user to analyze the data. NCSS gives control to users to select the layout, symbols, titles and many more.
- The end result obtained by NCSS can be directly used for further processing and is ready to be viewed, copied, pasted or saved.

14.7 Let us sum up:

This book has described how a number of important data technologies work and it has provided guidelines on the right way to use those technologies.

NOTES

We have covered technologies for storing data, accessing data, processing data, and publishing data on the web. We have also focused on writing computer languages.

14.8 Unit-End Exercise

1. List out few sector in which Statistical software is used?
2. Define SPSS.
3. What are the major features& benefits for using Wizard Mac?
4. Define Model Analyst's Toolkit (MAT)

14.8 Check your progress

1. What is statistical software?
2. What is SPSS software used for?
3. What are the different types of wizards?

14.9 Answer to check your progress

1. Statistical software is programs which are used for the statistical analysis of the collection, organization, analysis, interpretation and presentation of data.

2. SPSS is short for Statistical Package for the Social Sciences, and it's used by various kinds of researchers for complex statistical data analysis. The SPSS software package was created for the management and statistical analysis of social science data.

3. The three types of magic that mages can use are fire, frost, and arcane.

14.10 SUGGESTED READINGS

1. Dr Michael J de Smith (2018) Statistical Analysis Handbook, The Winchelsea Press, Drumlin Security Ltd, Edinburgh

2. R. Lyman Ott (2001) An Introduction to Statistical Methods and Data Analysis Fifth Edition, DUXBURY publication

NOTES

Further Readings

1. Bergeron B. (2003). Bioinformatics Computing - The Complete Practical Guide to Bioinformatics for Life Scientists, by Prentics- Hall, Inc., New Jersey 07458, USA, 1st edition, ISBN :81-203-2258-4.
2. David Mount, (2004), "Bioinformatics: Sequence and Genome Analysis"; Cold Spring harbor laboratory Press, US Revised Edition.
3. Baxevanis, A.D. and Francis Ouellellette, B.F. (2011) "Bioinformatics –a practical guide to the analysis of Genes and Proteins"; John Wiley & Sons, UK, Third Edition.
4. Scott Markel (2003) "Sequence Analysis in a Nutshell – A Guide to Common Tools & Databases"; O′Reilly; 1 edition, ISBN-13: 978-0596004941.
5. Zar, J.H. (1984) "Bio Statistical Methods"; Prentice Hall International Edition, USA
6. Gurumani,N., (2015)."An Introduction to Biostatistics", MJP Publisher, 2nd Edition
7. Warren, J., Gregory, E. and Grant, R. (2004) "Statistical methods in Bioinformatics"; First edition, Springer-Verlag, Berlin.
8. Milton, J.S. (1992) "Statistical methods in the Biological and Health Sciences"; Second Edition, McGraw Hill Publishers

Distance Education-CBCS-(2019-20 Academic Year Onwards)

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours

Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1. Define Bioinformatics
2. What is Operating System?
3. Comment on sequence analysis
4. Define and abbreviate BLAST
5. Explain Motif
6. Discuss on swiss prot
7. Define Biostatistics
8. Comment on mode
9. Explain regression
10. What is probability?

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a. writes a short note on servers and workstations

(or)

b. How to annotate and analyze genome sequences?

NOTES

12. a. Discuss the Basic Concepts of multiple sequence alignments

(OR)

b. Write in detail about Phylogenetic alignment

13. a. What are the concepts, terminology and theorems of probability?

(OR)

b. Write short note on measures of central tendencies

14. a. Discuss about ANOVA with suitable example

(OR)

b. Write an account of correlation types

15. a. Comment on various sampling techniques

(OR)

b. List out the importance of regression

Part – C (3 x 10 = 30 Marks)

Answer all questions

16. Give a detail account on UNIX and Linux operating system

17. Discuss more about the concept and procedure of protein modeling

18. Explain chi square rest with suitable example

19. Describe in detail about correlation and their methods

20. List out the statistical software in data analysis.

Common formula used in this chapter

NOTES

Unit-viii curriculam

1. **Mean**

$$\bar{x} = \frac{\sum_{I=1}^{n} x\, i}{N}$$

2. **Median** = value of item (n+1)/2

3. **Discrete variables**

$$S = \sqrt{\sum_{i=1}^{n}(Xi - \overline{X})^2}$$

UNIT IX - Curriculum

1. **Apriori probability:**

Apriori probability (P) = $\frac{\text{Number of favorable cases}}{\text{Total number of possible cases}}$

2. **Aposteriori probability:**

P = $\frac{\textbf{Number of times the event occurred}}{\textbf{Total number of trials}}$

3. **Binomial equation:**

$$(p+q)^n = p^n + {}_nC_{n-1}p^{n-1}q + {}_nC_{n-2}p^{n-2}q^2 + ... + {}_nC_r p^r q^{n-r} ... + {}_nC_1 pq^{n-1} + q^n$$

NOTES

4. Normal Distribution:

$$Y = \frac{N}{6\sqrt{2\pi}}.e^{\frac{-(X-\overline{X})^2}{26^2}}$$

UNIT X - Curriculum

1. Chi-Square Statistic:

$$X^2 = \sum\frac{(O-E)^2}{E}$$

2. P-Value:

$p2+ 2pq + q2 = 1$

3. Yate's correction factor

$$X^2 = \sum \frac{[(O-E)- 0.5]^2}{E}$$

Unit XI

1. Factorial design of experiment

1. $SS_{(total)} = \sum X^2 - \frac{(\sum X^2)}{N}$

NOTES

2. $SS_A = \frac{(\sum XA1)^2}{n\ A1} + \frac{(\sum XA2)}{n\ A2} + \frac{(\sum X)^2}{N}$

3. $SS_B = \frac{(\sum XB1)^2}{n\ B1} + \frac{(\sum XB2)}{n\ B2} + \frac{(\sum X)^2}{N}$

4. $SS_{\text{(interaction total)}} = \frac{(\sum XA1B1)}{n\ A1B1} + \frac{(\sum XA2B1)}{n\ A1B2} + \frac{(\sum XA2B2)}{n\ A2B2} - \frac{(\sum X)^2}{N}$

5. $SS_{\text{interaction}} = SS_{\text{interaction total}} - (SS_A + SS_B)$

6. $SS_{\text{Error}} = SS_{\text{total}} - SS_A + SS_B + SS_{\text{interaction}})$

2. F-test:

1. F = between group variability / within group variability.

3. Sum of Squares

1. $SS_{\text{total}} = \sum_{k=0}^{n} (\overline{X_J} - \bar{X})^2$

2. $SS_{\text{within}} = \sum_{j=1}^{k} \sum_{j=1}^{l} (X - \bar{X}\text{j})$

3. $SS_{\text{between}} = \sum_{j=1}^{k} (\overline{X_J} - \bar{X})^2$

4. $SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$.

4. The Degrees of Freedom (df)

1. $df_{\text{total}} = n - 1$

2. $df_{\text{within}} = k - 1$

NOTES

3. $df_{between} = n - k$

4. dft = dfw + dfb

5. The Mean Squares

1. $MS_{between} = \frac{SSbetween}{dfbetween}$

2. $MS_{within} = \frac{SSwithin}{dfwithin}$

6. Calculate the F Statistic

1. $F = \frac{MSbetween}{MSwithin}$

<u>Unit XII – Curriculum</u>

1. Karl Pearson's Coefficient of Correlation

1. $r = \frac{\sum(X - X)(Y - Y)}{nS_X S_Y}$

2. $r = \frac{\sum XY - \sum X \sum Y/n}{\sqrt{[\sum X^2 - (\sum X)^2][\sum Y^2 - (\sum Y)^2/n}}$

2. Spearman Rank Correlation

$r_R = 1 - \frac{6\sum idi^2}{n(n^2-1)}$

UNIT XIII - Curriculum

NOTES

1. Regression Equation

1. $\frac{(X - \bar{X}) = rS_X\,(Y - \bar{Y})}{S_Y}$

2. $\frac{(Y - \bar{Y}) = rS_X\,(X - \bar{X})}{S_Y}$

2. Regression Coefficient Constant Intercept of Regression Equation

1. $(X - \bar{X}) = \frac{S_X}{S_Y}(Y - \bar{Y})$

2. $A = \frac{S_X}{S_Y}$ or $A = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - (\sum Y)^2}$

3. $B = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{n\sum Y^2 - (\sum Y)^2}$

4. $A = \frac{rS_Y}{rS_X}$

5. $A = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$

6. $B = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$

DDE Course Material Preparation

M.Sc Microbiology

36443 Bioinformatics and Biostatistics

UNIT-I CURRICULAM

Structure

1.1 Introduction

1.2 Objectives

1.3 Biology in the computer age

1.3.1 How Is Computing Changing Biology?

1.3.2 The Eye of the Fly

1.3.3 Labels in Gene Sequences

1.3.4 The First Information Age in Biology

1.4 Computational Approaches to Biological Questions

1.4.1 Molecular Biology's Central Dogma

1.4.2 Replication of DNA

1.4.3 Genomes and Genes

1.4.4 Transcription of DNA

1.4.5 Translation of mRNA

1.4.6 Molecular Evolution

1.5 Basics of computers

1.5.1 CHARACTERISTICS OF COMPUTERS

1.6 Servers and workstation

1.6.1 Server Types and Services

1.6.2 Thin Servers

1.6.3 Thin Client Servers

1.7 Let Us Sum Up

1.8 Answers to check your progress

1.1 Introduction

Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modeling of data for analysis, visualization or prediction through the development of algorithms and software. In this unit describes the computational application of biology and needs. Thus, by studying the computers, it is very essential to know the relationship between computer with biology to the gain knowledge like servers, workstation, data storage, data retrieval, usage and characteristic of computers. Hence, it is essential to one who is willing to access the computers and servers.

1.2 Objectives

After going through the unit you will able to;

- Understand the concept of computer age
- Relate the meaning of computers with biology
- Arise the computational approaches to biological questions
- Identify the importance and key role of servers and workstation

1.3 Biology in the computer age

Biology is defined as the study of living thing to analyze the interaction of species and populations, to the function of tissues and cells within an individual organism. The computers in the biology which arise when the data are collected, stored, and interpret Now, at the beginning of the 21st century, we use sophisticated laboratory technology that allows us to collect data faster than we can interpret it. We have vast volumes of DNA sequence data at our fingertips. But how do we figure out which parts of that DNA control the various chemical processes of life? We know the function and structure of some proteins, but how do we determine the function of new proteins? And how do we

predict what a protein will look like, based on knowledge of its sequence? We understand the relatively simple code that translates DNA into protein. But how do we find meaningful new words in the code and add them to the DNA-protein dictionary? To overcome all these issues, Bioinformatics, a new emerging field arises to answer all the questions.

Bioinformatics is the science of using information to understand biology; it's the tool we can use to help us answer these questions and many others like them. Unfortunately, with all the hype about mapping the human genome, bioinformatics has achieved buzzword status; the term is being used in a number of ways, depending on who is using it. Strictly speaking, bioinformatics is a subset of the larger field of *computational biology,* the application of quantitative analytical techniques in modeling biological systems. In this book, we stray from bioinformatics into computational biology and back again. The distinctions between the two aren't important for our purpose here, which is to cover a range of tools and techniques we believe are critical for molecular biologists who want to understand and apply the basic computational tools that are available today. The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition. Researchers come to bioinformatics from many fields, including mathematics, computer science, and linguistics. Unfortunately, biology is a science of the specific as well as the general. Bioinformatics is full of pitfalls for those who look for patterns and make predictions without a complete understanding of where biological data comes from and what it means. By providing algorithms, databases, user interfaces, and statistical tools, bioinformatics makes it possible to do exciting things such as compare DNA sequences and generate results that are potentially significant. "Potentially significant" is perhaps the most important phrase. These new

tools also give you the opportunity to over-interpret data and assign meaning where none really exists. We can't overstate the importance of understanding the limitations of these tools. But once you gain that understanding and become an intelligent consumer of bioinformatics methods, the speed at which your research progresses can be truly amazing.

1.3.1 How Is Computing Changing Biology?

An organism's hereditary and functional information is stored as DNA, RNA, and proteins, all of which are linear chains composed of smaller molecules. These macromolecules are assembled from a fixed alphabet of well-understood chemicals: DNA is made up of four deoxyribonucleotides (adenine, thymine, cytosine, and guanine), RNA is made up from the four ribonucleotides (adenine, uracil, cytosine, and guanine), and proteins are made from the 20 amino acids. Because these macromolecules are linear chains of defined components, they can be represented as sequences of symbols. These sequences can then be compared to find similarities that suggest the molecules are related by form or function.

Sequence comparison is possibly the most useful computational tool to emerge for molecular biologists. The World Wide Web has made it possible for a single public database of genome sequence data to provide services through a uniform interface to a worldwide community of users. With a commonly used computer program called fsBLAST, a molecular biologist can compare an uncharacterized DNA sequence to the entire publicly held collection of DNA sequences. In the next section, we present an example of how sequence comparison using the BLAST program can help you gain insight into a real disease.

1.3.2 The Eye of the Fly

Fruit flies (*Drosophila melanogaster*) are a popular model system for the study of development of animals from embryo to adult. Fruit flies have a gene called *eyeless*, which, if it's "knocked out" (i.e., eliminated from the genome using molecular biology methods), results in fruit flies with no eyes. It's obvious that the *eyeless* gene plays a role in eye development.

Researchers have identified a human gene responsible for a condition called *aniridia*. In humans who are missing this gene (or in whom the gene has mutated just enough for its protein product to stop functioning properly), the eyes develop without irises. If the gene for *aniridia* is inserted into an eyeless drosophila "knock out," it causes the production of normal drosophila eyes. It's an interesting coincidence. Could there be some similarity in how *eyeless* and *aniridia* function, even though flies and humans are vastly different organisms? Possibly. To gain insight into how *eyeless* and *aniridia* work together, we can compare their sequences. Always bear in mind, however, that genes have complex effects on one another. Careful experimentation is required to get a more definitive answer.

As little as 15 years ago, looking for similarities between *eyeless* and *aniridia* DNA sequences would have been like looking for a needle in a haystack. Most scientists compared the respective gene sequences by hand-aligning them one under the other in a word processor and looking for matches character by character. This was time-consuming, not to mention hard on the eyes.

In the late 1980s, fast computer programs for comparing sequences changed molecular biology forever. Pairwise comparison of biological sequences is the foundation of most widely used bioinformatics techniques. Many tools that are widely

available to the biology community-including everything from multiple alignment, phylogenetic analysis, motif identification, and homology-modeling software, to web-based database search services-rely on pairwise sequence-comparison algorithms as a core element of their function.

1.3.3 Labels in Gene Sequences

Before you rush off to compare the sequences of *eyeless* and *aniridia* with BLAST, let us explore a little bit about how sequence alignment works.

It's important to remember that biological sequence (DNA or protein) has a chemical function, but when it's reduced to a single-letter code, it also functions as a unique label, almost like a bar code. From the information technology point of view, sequence information is priceless. The sequence label can be applied to a gene, its product, its function, its role in cellular metabolism, and so on. The user searching for information related to a particular gene can then use rapid pairwise sequence comparison to access any information that's been linked to that sequence label.

The most important thing about these sequence labels, though, is that they don't just uniquely identify a particular gene; they also contain biologically meaningful patterns that allow users to compare different labels, connect information, and make inferences. So not only can the labels connect all the information about one gene, they can help users connect information about genes that are slightly or even dramatically different in sequence.

If simple labels were all that was needed to make sense of biological data, you could just slap a unique number (e.g., a GenBank ID) onto every DNA sequence and be done with it. But biological sequences are related by evolution, so a partial pattern match between two sequence labels is a significant find. BLAST differs from simple keyword

searching in its ability to detect partial matches along the entire length of a protein sequence.

BLAST finds local regions that match even in pairs of sequences that aren't exactly the same overall. It extends matches beyond a single-character difference in the sequence, and it keeps trying to extend them in all directions until the overall score of the sequence match gets too small. As a result, BLAST can detect patterns that are imperfectly replicated from sequence to sequence, and hence distant relationships that are inexact but still biologically meaningful.

Depending on the quality of the match between two labels, you can transfer the information attached to one label to the other. A high-quality sequence match between two full-length sequences may suggest the hypothesis that their functions are similar, although it's important to remember that the identification is only tentative until it's been experimentally verified. In the case of the *eyeless* and *aniridia* genes, scientists hope that studying the role of the *eyeless* gene in Drosophila eye development will help us understand how *aniridia* works in human eye development.

1.3.4 The First Information Age in Biology

Biology as a science of the specific means that biologists need to remember a lot of details as well as general principles. Biologists have been dealing with problems of information management since the 17th century.

The roots of the concept of evolution lie in the work of early biologists who catalogued and compared species of living things. The cataloguing of species was the preoccupation of biologists for nearly three centuries, beginning with animals and plants and continuing with microscopic life upon the invention of the compound microscope. New forms of life and fossils of previously unknown, extinct life forms are still being

discovered even today.

All this cataloguing of plants and animals resulted in what seemed a vast amount of information at the time. In the mid-16th century, Otto Brunfels published the first major modern work describing plant species, the *Herbarium vitae eicones*. As Europeans traveled more widely around the world, the number of catalogued species increased, and botanical gardens and herbaria were established. The number of catalogued plant types was 500 at the time of Theophrastus, a student of Aristotle. By 1623, Casper Bauhin had observed 6,000 types of plants. Not long after John Ray introduced the concept of distinct species of animals and plants, and developed guidelines based on anatomical features for distinguishing conclusively between species. In the 1730s, Carolus Linnaeus catalogued 18,000 plant species and over 4,000 species of animals, and established the basis for the modern taxonomic naming system of kingdoms, classes, genera, and species. By the end of the 18th century, Baron Cuvier had listed over 50,000 species of plants.

It was no coincidence that a concurrent preoccupation of biologists, at this time of exploration and cataloguing, was classification of species into an orderly taxonomy. A botany text might encompass several volumes of data, in the form of painstaking illustrations and descriptions of each species encountered. Biologists were faced with the problem of how to organize, access, and sensibly add to this information. It was apparent to the casual observer that some living things were more closely related than others. A rat and a mouse were clearly more similar to each other than a mouse and a dog. But how would a biologist know that a rat was like a mouse (but that rat was not just another name for mouse) without carrying around his several volumes of drawings? A nomenclature that uniquely identified each living thing and summed up its presumed relationship with

other living things, all in a few words, needed to be invented.

The solution was relatively simple, but at the time, a great innovation. Species were to be named with a series of one-word names of increasing specificity. First a very general division was specified: animal or plant? This was the kingdom to which the organism belonged. Then, with increasing specificity, came the names for class, genera, and species. This schematic way of classifying species, as illustrated in Figure 1.1, is now known as the "Tree of Life."

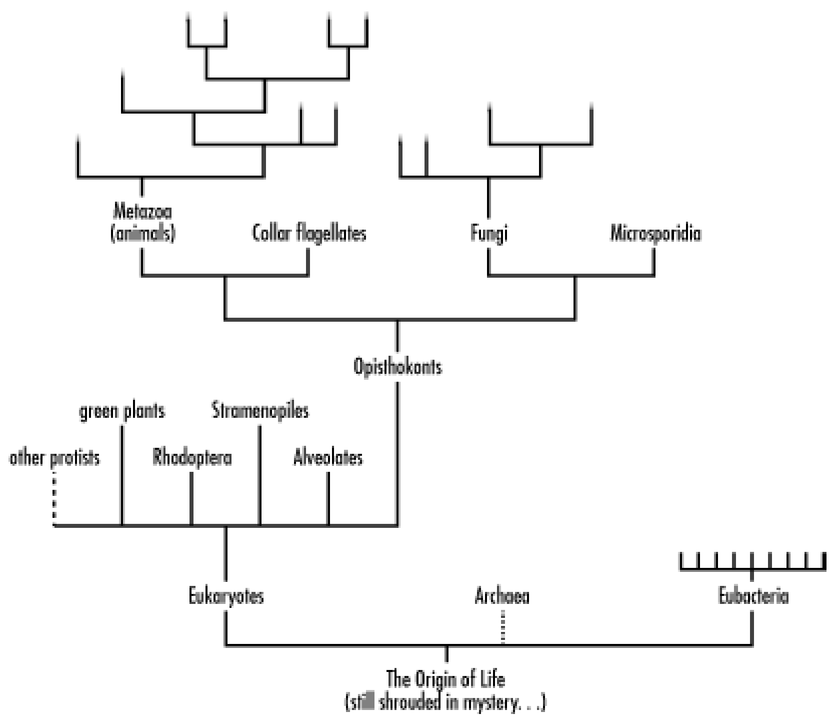


Figure 1.1. The "Tree of Life" represents the nomenclature system that classifies species

A modern taxonomy of the earth's millions of species is too complicated for even the most zealous biologist to memorize, and fortunately computers now provide a way to maintain and access the taxonomy of species. The University of Arizona's Tree of Life project and NCBI's Taxonomy database are two examples of online taxonomy projects. Taxonomy was the first informatics problem in biology. Now, biologists have reached a similar point of information overload by collecting and cataloguing information about individual genes. The problem of organizing this information and sharing knowledge with the scientific community at the gene level isn't being tackled by developing a nomenclature. It's being attacked directly with computers and databases from the start.

The evolution of computers over the last half-century has fortuitously paralleled the developments in the physical sciences that allow us to see biological systems in increasingly fine detail. Figure 1.2 illustrates the astonishing rate at which biological knowledge has expanded in the last 20 years.

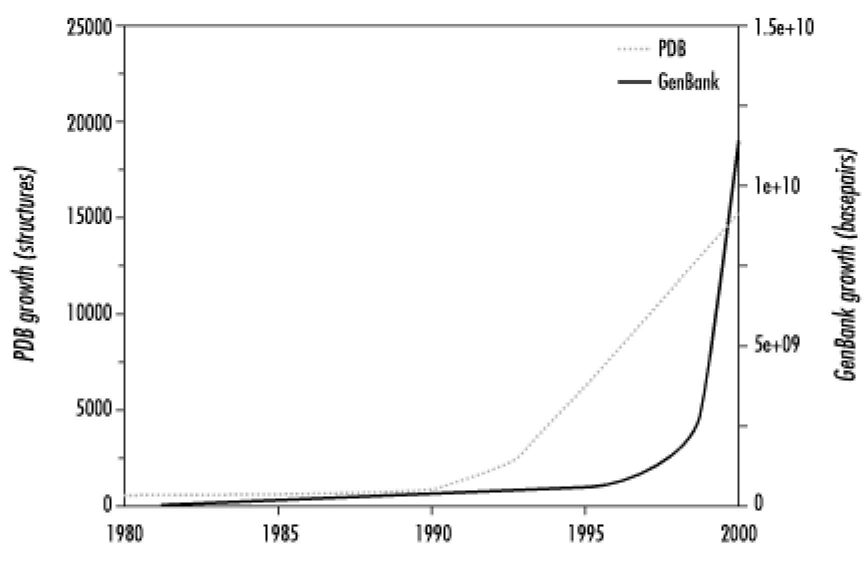


Figure 1.2. The growth of GenBank and the Protein Data Bank

Simply finding the right needles in the haystack of information that is now available can be a research problem in itself. Even in the late 1980s, finding a match in a sequence database was worth a five-page publication. Now this procedure is routine, but there are many other questions that follow on our ability to search sequence and structure databases. These questions are the impetus for the field of bioinformatics.

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. What is Bioinformatics?

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

1.4 Computational Approaches to Biological Questions

There is a standard range of techniques that are taught in bioinformatics courses. Currently, most of the important techniques are based on one key principle: that sequence and structural homology (or similarity) between molecules can be used to infer structural and functional similarity. In this unit, we will give you an overview of the standard computer techniques available to biologists and we will discuss how specific software packages implement these techniques and how you should use them.

1.4.1 Molecular Biology's Central Dogma

Before we go any further, it's essential that you understand some basics of cell and molecular biology. If you're already familiar with DNA and protein structure, genes, and the processes of transcription and translation, feel free to skip ahead to the next section.

The central dogma of molecular biology states that:

> *DNA acts as a template to replicate itself, DNA is also transcribed into RNA, and RNA is translated into protein.*

As you can see, the central dogma sums up the function of the genome in terms of information. Genetic information is conserved and passed on to progeny through the process of replication. Genetic information is also used by the individual organism through the processes of transcription and translation. There are many layers of function, at the structural, biochemical, and cellular levels, built on top of genomic information. But in the end, all of life's functions come back to the information content of the genome.

Put another way, genomic DNA contains the master plan for a living thing. Without DNA, organisms wouldn't be able to replicate themselves. The raw "one-dimensional" sequence of DNA, however, doesn't actually do anything biochemically; it's only information, a blueprint if you will, that's read by the cell's protein synthesizing machinery. DNA sequences are the punch cards; cells are the computers.

DNA is a linear polymer made up of individual chemical units called *nucleotides* or *bases*. The four nucleotides that make up the DNA sequences of living things (on Earth, at least) are adenine, guanine, cytosine, and thymine—designated A, G, C, and T, respectively. The order of the nucleotides in the linear DNA sequence contains the instructions that build an organism. Those instructions are read in processes called replication, transcription, and translation.

1.4.2 Replication of DNA

The unusual structure of DNA molecules gives DNA special properties. These properties allow the information stored in DNA to be preserved and passed from one cell to another, and thus from parents to their offspring. Two molecules of DNA form a double-helical structure, twining around each other in a regular pattern along their full length—which can be millions of nucleotides. The halves of the double helix are held together by bonds between the nucleotides on each strand. The nucleotides also bond in particular ways: A can pair only with T, and G can pair only with C. Each of these pairs is referred to as a *base pair,* and the length of a DNA sequence is often described in base pairs (or bp), kilobases (1,000 bp), megabases (1 million bp), etc.

Each strand in the DNA double helix is a chemical "mirror image" of the other. If there is an A on one strand, there will always be a T opposite it on the other. If there is a C on one strand, its partner will always be a G.

When a cell divides to form two new *daughter cells*, DNA is *replicated* by untwisting the two strands of the double helix and using each strand as a template to build its chemical mirror image, or *complementary strand*. This process is illustrated in Figure 1.3.

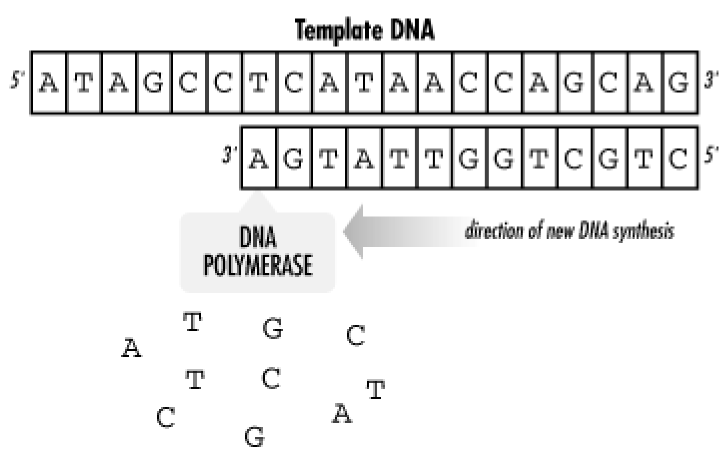


Figure 1.3. Schematic replication of one strand of the DNA helix

1.4.3 Genomes and Genes

The entire DNA sequence that codes for a living thing is called its *genome*. The genome doesn't function as one long sequence, however. It's divided into individual genes. A *gene* is a small, defined section of the entire genomic sequence, and each gene has a specific, unique purpose.

There are three classes of genes. *Protein-coding* genes are templates for generating molecules called proteins. Each *protein* encoded by the genome is a chemical machine with a distinct purpose in the organism. *RNA-specifying* genes are also templates for chemical machines, but the building blocks of RNA machines are different from those that make up proteins. Finally, *untranscribed genes* are regions of genomic DNA that have some functional purpose but don't achieve that purpose by being transcribed or translated to create another molecule.

1.4.4 Transcription of DNA

DNA can act not only as a template for making copies of itself but also as a blueprint for a molecule called ribonucleic acid (RNA). The process by which DNA is transcribed into RNA is called *transcription* and is illustrated in Figure 1.4. RNA is structurally similar to DNA. It's a polymeric molecule made up of individual chemical units, but the chemical backbone that holds these units together is slightly different from the backbone of DNA, allowing RNA to exist in a single-stranded form as well as in a double helix. These single-stranded molecules still form base pairs between different parts of the chain, causing RNA to fold into 3D structures. The individual chemical units of RNA are designated A, C, G, and U (uracil, which takes the place of thymine).

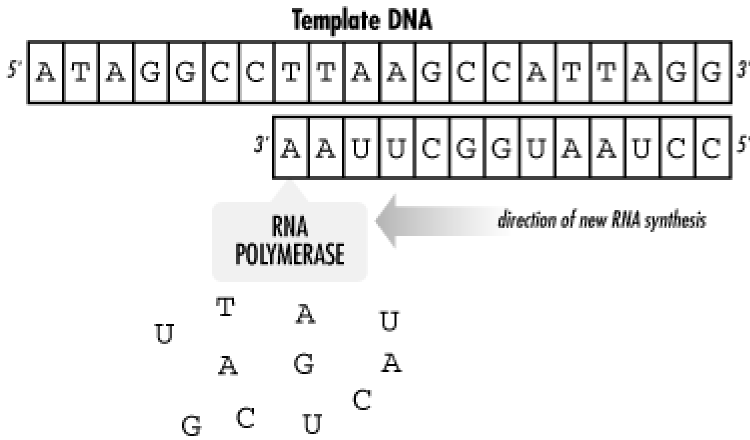


Figure 1.4. Schematic of DNA being transcribed into RNA

The genome provides a template for the synthesis of a variety of RNA molecules: the three main types of RNA are messenger RNA, transfer RNA, and ribosomal RNA. *Messenger* RNA (mRNA) molecules are RNA transcripts of genes. They carry information from the genome to the ribosome, the cell's protein synthesis apparatus. *Transfer* RNA (tRNA) molecules are untranslated RNA molecules that transport amino acids, the building blocks of proteins, to the ribosome. Finally, *ribosomal* RNA (rRNA) molecules are the untranslated RNA components of ribosomes, which are complexes of protein and RNA. rRNAs are involved in anchoring the mRNA molecule and catalyzing some steps in the translation process. Some viruses also use RNA instead of DNA as their genetic material.

1.4.5 Translation of mRNA

Translation of mRNA into protein is the final major step in putting the information in the genome to work in the cell. Like DNA, proteins are linear polymers built from an alphabet of chemically variable units. The protein alphabet is a set of small molecules called *amino acids*.

Unlike DNA, the chemical sequence of a protein has physicochemical "content" as

well as information content. Each of the 20 amino acids commonly found in proteins has a different chemical nature, determined by its side chain—a chemical group that varies from amino acid to amino acid. The chemical sequence of the protein is called its *primary structure*, but the way the sequence folds up to form a compact molecule is as important to the function of the protein as is its primary structure. The secondary and tertiary structure elements that make up the protein's final fold can bring distant parts of the chemical sequence of the protein together to form functional sites.

As shown in Figure 1.5, the *genetic code* is the code that translates DNA into protein. It takes three bases of DNA (called a *codon*) to code for each amino acid in a protein sequence. Simple combinatorics tells us that there are 64 ways to choose 3 nucleotides from a set of 4, so there are 64 possible codons and only 20 amino acids. Some codons are redundant; others have the special function of telling the cell's translation machinery to stop translating an mRNA molecule. Figure 1.6 shows how RNA is translated into protein.

First Position	Second Position: U		C		A		G		Third Position
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met (start)	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 1.5. The genetic code

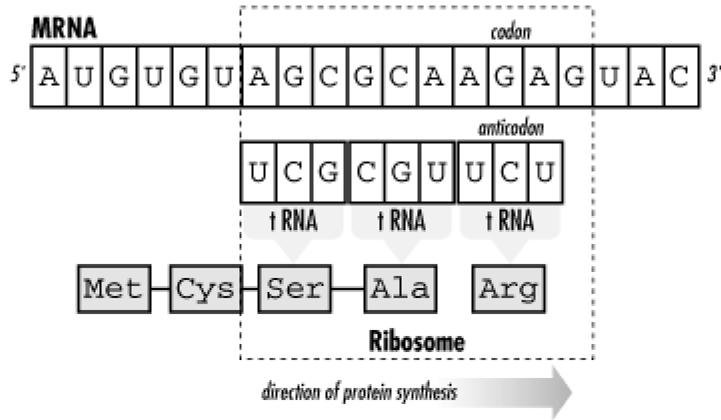


Figure 1.6. Synthesis of protein with standard base pairing

1.4.6 Molecular Evolution

Errors in replication and transcription of DNA are relatively common. If these errors occur in the reproductive cells of an organism, they can be passed to its progeny. Alterations in the sequence of DNA are known as *mutations*. Mutations can have harmful results—results that make the progeny less likely to survive to adulthood. They can also have beneficial results, or they can be neutral. If a mutation doesn't kill the organism before it reproduces, the mutation can become fixed in the population over many generations. The slow accumulation of such changes is responsible for the process known as *evolution*. Access to DNA sequences gives us access to a more precise understanding of evolution. Our understanding of the molecular mechanism of evolution as a gradual process of accumulating DNA sequence mutations is the justification for developing hypotheses based on DNA and protein sequence comparison.

Check your progress-2

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Comment on computational approaches to biological question.

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

1.5 Basics of computers

A computer is a programmable machine designed to perform arithmetic and logical operations automatically and sequentially on the input given by the user and gives the desired output after processing. Computer components are divided into two major categories namely hardware and software. Hardware is the machine itself and its connected devices such as monitor, keyboard, mouse etc. Software are the set of programs that make use of hardware for performing various functions.

1.5.1 CHARACTERISTICS OF COMPUTERS

Speed

Computers work at an incredible speed. A powerful computer is capable of performing about 3-4 million simple instructions per second.

Accuracy

In addition to being fast, computers are also accurate. Errors that may occur can almost always be attributed to human error (inaccurate data, poorly designed system or faulty instructions/programs written by the programmer)

Diligence

Unlike human beings, computers are highly consistent. They do not suffer from human traits of boredom and tiredness resulting in lack of concentration. Computers, therefore, are better than human beings in performing voluminous and repetitive jobs.

Versatility

Computers are versatile machines and are capable of performing any task as long as it can be broken down into a series of logical steps. The presence of computers can be seen in almost every sphere – Railway/Air reservation, Banks, Hotels, Weather forecasting and many more.

Storage Capacity

Today's computers can store large volumes of data. A piece of information once recorded (or stored) in the computer, can never be forgotten and can be retrieved almost instantaneously.

1.6 Servers and workstation

A complete understanding of server hardware can take a great deal of study. This chapter provides an overview of the server and identifies different server types and their roles. You will be introduced to some of the hardware that makes the server unique from the ordinary PC, and you will learn about RAID and other storage systems.

1.6.1 Server Types and Services

`Servers provide a variety of services. Some of the services a server can provide are authentication and security, Web, mail, and print. A server can be called by many names. For example, it can be called an authentication and security database server, Web server, mail server, and print server, Figure 9-1. A network may have a single server that provides a variety of services, or it may have a group of servers, each providing a specifi c service. A small network usually has one server set up to handle many different services. A large network usually has several servers, each providing a different service or set of services. For example, a large

corporation may use one server to handle e-mail requests and Web hosting, another server to serve as a domain controller to provide security for the entire network, and another server to provide application software, a database for its clients, and support for print operations. The more services a server provides and the more clients it services creates a higher demand on the server. A single server with a limited amount of services may be fine for a relatively small number of users, typically less than 25. A commercial enterprise spanning across the country or world may require hundreds of servers. The network administrator or network designer decides the number and capacity of individual servers needed by taking into account the number of users and the systems predicted network traffic. Each network system is uniquely designed, even though each network has many similarities. Some network equipment providers have software programs that help you design a network. You simply enter information, such as the number of clients, offices, cities, and countries and the type of software and services to be provided. After all the information is collected, the software program provides an estimate of the size and number of servers required.

1.6.2 Thin Servers

A thin server is a server that has only the hardware and software needed to support and run a specifi c function, such as Web services, print services, and file services. It is more economical to use a thin server as a print server than to tie up a more expensive server simply to handle printing on a network. IBM markets a thin server, which consists of a sealed box that contains only the essential hardware and software required for supporting the server's dedicated function.

1.6.3 Thin Client Servers

A thin client server is a server that provides applications and processing power to a thin client. Thin client servers run terminal server software and may have more RAM and hard disk drive storage than needed. A thin client is a computer that typically has a minimal amount of processing power and memory. It relies on the server's processor and memory for processing data and running software applications. Thin clients are becoming common in industry for such applications as hotel bookings, airline ticketing, and medical record access. When budgets are tight, you can install what is normally considered an obsolete PC on a thin client network. The obsolete PC is obsolete because of its processing speed, lack of storage, and inability to run new software. The thin client server can provide all the services that are required by the obsolete PC, thus making the obsolete PC a useful workstation. Do not confuse a thin client with a dumb terminal. A dumb terminal sends user input to a mainframe. Dumb terminals have absolutely no computing power, operating system, hard disk drive, BIOS, and RAM. A thin client may not need a hard disk drive or an operating system; however, it is still a full-fledged computer because it has a CPU and processing power. Windows 2000 Server, Windows Server 2003, and Windows Server 2008 come with Terminal Services software. Once Terminal Services is set up on the server, you have the option to generate a disk that will assist you in automatically setting up thin client workstations.

Check your progress-3

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Write the characteristics of computers

…………………………………………………………………………………………………

ii. Give a note on different types of servers

………………………………………………………………………………………………..

1.7 LET US SUM UP

In this unit, you have learnt about the computer age, computational approaches to biological questions and characteristics of computers. This knowledge would make understand what is computer and how it plays major role biology. The concept such as computer age has given detail knowledge about how computers interpret and evolution of biology. The application of computer age would pave the key attention to the student. The servers and workstation gave a how the networks are interlinked and the main access worldwide.

1.8 ANSWERS TO CHECK YOUR PROGRESS

1**.** Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modeling of data for analysis, visualization or prediction through the development of algorithms and software.

2. Biological software's information, Protein structure identification, structural and functional annotation, how to retrieve the data and etc.

3. Speed, Accuracy, Diligence, Versatility and Storage Capacity

4. Thin and Thin Client servers.

<u>Distance Education-CBCS-(2019-20 Academic Year Onwards)</u>

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.

UNIT-II CURRICULAM

Structure

1.1 Introduction

An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer. In this unit, we cover the software's related information to your study. The operating system which gives details knowledge about the computer progress. The basic characteristics and working mechanism of UNIX and LINUX will be studied in detail. Hence the unit would pave the attention to the student for working the computers like access operating system, search engine world wide web and finding and collecting article through their study.

1.2 Objectives

After going through the unit you will able to;

- Understand the concept of operating system
- Know about UNIX and Linux operating system and their evolution.
- To identify the scientific articles and their related databases
- The data encrypting and detail difference network analysis will be studied

1.3 Operating systems

An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer. There are many OS were available in the market such as Linux, Windows, Mac OS, Solaris and etc. The above

mentioned OS is freely and commercially available. In our unit we cover UNIX and Linux OS. These two OS is freely available and graphical user interface (GUI) system.

1.3.1 UNIX OPERATING SYSTEM

UNIX is a powerful operating system for multiuser computer systems. It has been in existence for over 25 years, and during that time has been used primarily in industry and academia, where networked systems and multiuser high-performance computer systems are required. UNIX is optimized for tasks that are only fairly recent additions to personal-computer operating systems, or which are still not even available in some PC operating systems: networking with other computers, initiating multiple asynchronous tasks, retaining unique information about the work environments of multiple users, and protecting the information stored by individual users from other users of the system. Unix is the operating system of the World Wide Web; the software that powers the Web was invented in Unix, and many if not most web servers run on Unix servers.

UNIX is rich in commands and possibilities. Every distribution of UNIX comes with a powerful set of built-in programs. Everything from networking software to word-processing software to electronic mail and news readers is already a part of UNIX. Many other programs can be downloaded and installed on UNIX systems for free.

It might seem that there's far too much to learn to make working on a UNIX system practical. It's possible, however, to learn a subset of UNIX and to become a productive Unix user without knowing or using every program and feature.

1.3.2 UNIX Shell

When you log into a UNIX system or open a new window in your system's window manager interface, the system automatically starts a program called a shell for you. The shell program interprets the commands you enter and provides you with a working environment and an interface to the operating system. It's possible to work in UNIX without the shell using graphical file manager tools, but you'll find that many shell commands are useful for data processing and analysis.

1.3.3 Standard Input and Output

By default, many UNIX commands read from standard input and send their output to standard output. Standard input and output are file descriptors associated with your terminal. A program reading from standard input will simply hang out and wait for you to type something on your keyboard and press the Enter key. A program writing to standard output spews its output to your terminal, sometimes far faster than you can read it.

Some UNIX commands read a hyphen (-) surrounded by whitespace on either side as "data from standard input." This construct can then be used in place of a filename in the command line. Absence of an output filename is sufficient to cause the program to write to standard output.

1.3.4 Graphical Interfaces for UNIX

Although UNIX is a text-based operating system, you no longer have to experience it as a black screen full of glowing green or amber letters. Most UNIX systems use a variant of the X Window System. The X Window System formats the screen environment and

allows you to have multiple windows and applications open simultaneously. X windows are customizable so that you can use menu bars and other widgets much like PC operating systems. Individual UNIX shells on the host machine as well as on networked machines are opened as windows, allowing you to exploit Unix's multitasking capabilities and to have many shells active simultaneously. In addition to UNIX shells and tools, there are many applications that take advantage of the X system and use X windows as part of their graphical user interfaces, allowing these applications to be run while still giving access to the UNIX command line.

The GNOME and KDE desktop environments, which are included in most major Linux distributions, make your Linux system look even more like a personal computer. Toolbars, visual file managers, and a configurable desktop replicate the feeling of a Windows or Mac work environment, except that you can also open a shell window and run UNIX programs.

1.3.5 Linux

Linux (LIH-nucks) is an open source version of Unix, named for its original developer, Linus Torvalds of the University of Helsinki in Finland. Originally undertaken as a one-man project to create a free UNIX for personal computers, Linux has grown from a hobbyist project into a product that, for the first time, gives the average personal-computer user access to a UNIX system.

In this unit, we focus on Linux for three reasons. First, with the availability of Linux, Unix is cheap (or free, if you have the patience to download and install it). Second, under Linux, inexpensive PCs regarded as "obsolete" by Windows users become startlingly flexible and useful workstations. The Linux operating system can be configured to use a

much smaller amount of system resources than the personal computer operating systems, which means computers that have been outgrown by the ever-expanding system requirements of PC programs and operating systems, can be given a new lease on life by being reconfigured to run Linux. Third, Linux is an excellent platform for developing software, so there's a rich library of tools available for computational biology and for research in general.

1.3.6 Installing the software in Linux

1. Download the source code distribution. Use binary mode; compressed text files are encoded.
2. Download and read the instructions (usually a README or INSTALL file; sometimes you have to find it after you extract the archive).
3. Make a new directory and move the archive into it, if necessary.
4. uncompress (*.Z) or gunzip (*.gz) the file.
5. Extract the archive using tar xvf or as instructed.
6. Follow the instructions (did we say that already?).
7. Run the configuration script, if present.
8. Run make if a Makefile is present.
9. If a Makefile isn't present and all you see are *.f or *.c files, use gcc or g77 to compile them, as discussed earlier.
10. Run the installation script, if present.
11. Link the newly created binary executable into one of the binary-containing directories in your path using ln -s (this is usually part of the previous step, but if there is no installation script, you may need to create the link by hand).

1.3.7 Paths to Files and Directories

Each file on the filesystem can be uniquely identified by a combination of a filename and a path. You can reference any file on the system by giving its full name, which begins with a / indicating the root directory, continues through a list of subdirectories (the components of the path) and ends with the filename. The full name, or *absolute path*, of a file in someone's home directory might look like this:

/home/jambeck/mustelidae/weasels.txt

1.3.8 Applications of Linux

Linux is a free available OS for deskotop and laptop platform with GUI constructing interface. Many popular applications such as Mozilla Firefox, OpenOffice.org/LibreOffice and Blender has available in Linux OS. Besides externally visible components, such as X window managers, a non-obvious but quite central role is played by the programs hosted by freedesktop.org, such as D-Bus or PulseAudio; both major desktop environments (GNOME and KDE) include them, each offering graphical front-ends written using the corresponding toolkit (GTK+ or Qt). A display server is another component, which for the longest time has been communicating in the X11 display server protocol with its clients; prominent software talking X11 includes the X.Org Server and Xlib. Linux distributions have long been used as server operating systems.

Check your progress-1

Note: a. Write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. What is operating system?

…………………………………………………………………………………………………

1.4 World-Wide Web

The World-Wide Web is a collection of documents and services, distributed across the Internet and linked together by hypertext links. The web is therefore a subset of the Internet, not the same thing. Much of what is available on the web consists of web documents, which are often (somewhat misleadingly) called “home pages.”

A home page appears to be a single entity, but is actually made up of a number of separate and distinct files, which may incorporate one or more of the following, in different configurations: text

*graphics

*animation

*audio

*video

*hyperlinks (text or graphics which lead you from document to document)

*interactive element, including: specially embedded programs such as: * plug-ins (downloadable sets of software that enable the user to use part of a web document.)

The base document of a home page is a file which is mostly text, containing commands (“markup”) which determines how the above elements are configured. The rules governing this markup form a simple programming language called “HTML”. HTML stands for “Hypertext Markup Language”. As mentioned above, the Web also provides access to services. Some of these services are assembled “on the fly” by a computer program that accesses information available in some other form, and presents the information in the same format as any other web

page. The Web is therefore a user interface providing access to many types of information available on the Internet.

1.5 Search Engines

A search engine is software, usually accessed on the Internet that searches a database of information according to the user's query. The engine provides a list of results that best match what the user is trying to find. Today, there are many different search engines available on the Internet, each with their own abilities and features. The first search engine ever developed is considered Archie, which was used to search for FTP files and the first text-based search engine is considered Veronica. Currently, the most popular and well-known search engine is Google. Other popular search engines include AOL, Ask.com, Baidu, Bing, and Yahoo.

1.5.1 Search Engine Working

Web crawler, database and the search interface are the major component of a search engine that actually makes search engine to work. Search engines make use of Boolean expression AND, OR, NOT to restrict and widen the results of a search. Following are the steps that are performed by the search engine:

- The search engine looks for the keyword in the index for predefined database instead of going directly to the web to search for the keyword.
- It then uses software to search for the information in the database. This software component is known as web crawler.
- Once web crawler finds the pages, the search engine then shows the relevant web pages as a result. These retrieved web pages generally include title of page, size of text portion, first several sentences etc.

- User can click on any of the search results to open it.

Check your progress-2

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. What is World Wide Web and note the different search engines

…………………………………………………………………………………………………

1.6 Pubmed

PubMed is the U.S. National Library of Medicine’s (NLM) premiere search system for health information.

1.6.1 PubMed Content

Nearly 25 million citations including:

• Publisher supplied citations that will be analyzed to receive full indexing for MEDLINE if they are biomedical in nature

• In-process citations that have not yet been analyzed and indexed for MEDLINE

• Indexed for MEDLINE citations of articles from about 5600 regularly indexed journals; MEDLINE makes up nearly 90% of PubMed.

1.6.2 PubMed Features

• Sophisticated search capabilities, including spell checker, Advanced Search Builder, and special tools for searching for clinical topics

• Assistance in finding search terms using the MeSH (Medical Subject Heading) database of MEDLINE's controlled vocabulary

• Ability to store citation collections and to receive email updates from saved searches using PubMed's My NCBI

• Links to full-text articles, to information about library holdings, and to other NLM databases and search interfaces

1.6.3 Clinical Queries

PubMed Clinical Queries makes it easy to find articles that report applied clinical research. Click on the link from the PubMed homepage, then enter a search term in the box. Click the Search button. Click See all at the bottom of the page to return to PubMed.

Clinical Study Categories displays results by diagnosis, etiology, therapy, etc. Use the dropdown menus to change the category or scope.

Systematic Reviews displays evidence-based medicine citations including systematic reviews, meta-analyses, and guidelines.

Medical Genetics displays citations focused on diagnosis, management, genetic counseling, and related topics. Select All or a specific topic from the drop-down menu.

1.7 Public Biological databases

The past few decades computational biology has become an emerging technology and widespread applications of computers are used for analysis and modeling of biological data.

Storage, retrieval and analysis of biological dataset maintained by centralized resources worldwide are the major aspect of this revolution. Dramatic increases have been observed in genomic and proteomic data as a resultant of advancement in molecular biology techniques and high throughput methods. These methods include a wide range of information such as genetic and physical genomic maps, polymorphisms, biobliographic information, molecular chemical properties and macromolecular sequences and structures etc., Concurrently, the fast development of biomedical knowledge, reduction in computer costs, wide spread of internet access, and the recent emergence in the field of high throughput structural and functional genomic technologies has directed to a rapid expansion of electronically available data. Such data has deposited in public archives as a various biological databases and can be used by the scientific community for querying and retrieving of necessary information. The Molecular Biology Database Collection by Micheal Galperin is a collection of several databases. The databases included in the collection that are freely available. The recent update includes 719 databases. The databases are arranged in a hierarchical classification and categorized into three different branches. National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL), and DNA Data Bank of Japan (DDBJ) are the major organizations whereas a majority of the tools and databases are being maintained.

Early in the 1960s and 1970s, Margaret Dayhoff was the first one who worked on protein sequences and was first assembled and made freely available. Now, this database named as the international Protein Information Resources (PIR). Later, SWISS-PROT sequence databases was framed and publicly announced in 1980 (Bairoch and Boeckmann, 1991). Concomitantly,

various databases were initiated to flourish in order to handle the increasing quantities of sequence data being generated worldwide.

Generally, databases can be classified into primary, secondary, and composite databases. Primary database contain the information that are directly deposited by the submitter. NCBI, EMBL, DDBJ, SWISS-PROT and PIR are the widely used primary databases for the sequences (nucleotide and proteins) and Protein Data Bank (PDB) have been universally used for protein three dimensional structures. Secondary database includes information such as conserved sequence, signature sequences, conserved secondary structure motifs and active site residues of the protein families using multiple sequence alignment. Various researchers at different individual laboratories created and maintained different databases such as SCOP, CATH, PROSITE, and eMOTIF , CDART, and protein interactions.

1.7.1 Genomic Sequence Databases – GenBank, EMBL, DDBJ

Genbank, was built and maintained by the National Center for Biotechnology Information (NCBI). It is part of the International Nucleotide Sequence Database in collaboration with the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences , acquire initially through submissions from individual laboratories and batch submissions from large-scale sequencing projects. NCBI’s retrieval system helps to access GenBank and Entrez used to combine information from the major DNA and protein sequence databases along with mapping, taxonomy genome, protein structure and domain information, and the literature information through PubMed. The NCBI has integrated server called BLAST (Basic Local Alignment Searching Algorithm) that

identifies similar sequences for the query sequence. Up-to-now nearly 60 million sequences are available and it's growing aggressively. Genbank also provides the FTP access to the user wher they can retrieve the large set of data for their analysis from ftp://ftp.ncbi.nih.gov. Each sequence in GenBank is annotated with literature reference, size, sequence features, organism, and protein translations. The GenBank sequences are arranged under several divisions viz. EST, ENV, Plant sequences (PLN), Genome survey sequences (GSS), rodents (ROD), Sequence tagged site (STS) etc.,

GenBank entry includes a brief description of the sequence, the scientific name, and taxonomy of the source organism, bibliographic references, and a table of features. Table I.1. list major sequence databases and Table I.2. Lists different sequence divisions in Genbank.

1.7.2 Protein Sequence Databases

Swiss Protein Databank (SWISS-PROT) (Boeckmann et al., 2003), the translation of the DNA sequences in EMBL (TrEMBL), Protein Information Resource (PIR), the Munich Information Center for Proteins (MIPS), and the 3D structures in the Protein Data Bank (PDB). Protein databases are growing faster than DNA databases.

Check your progress-3

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Comment on the clinical queries of pubmed and different biological databases.

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

1.8 LET US SUM UP

In this unit, you have learnt about the operating system, World Wide Web, search engines finding scientific articles. The operating system gave detail information about how our computer works and how to change our input into output. The detail application of Linux will helpful to your study like writing algorithms, developing database and command prompt methods. The World Wide Web and search engines gave the internet access and finding related articles. The database information is helpful to retrieve and store the data.

1.9 ANSWERS TO CHECK YOUR PROGRESS

1**.** An Operating system (OS) is software which acts as an interface between the end user and computer hardware. Every computer must have at least one OS to run other programs. An application likes Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task. The OS helps you to communicate with the computer.

2. The World-Wide Web is a collection of documents and services, distributed across the Internet and linked together by hypertext links. Search Engines: Google, AOL, Ask.com, Baidu, Bing, and Yahoo

3. Clinical Queries: PubMed Clinical Queries, Clinical Study Categories, Systematic Reviews and Medical Genetics

Biological databases: GenBank, EMBL, DDBJ, SWISS-PROT, PIR and etc..

<u>Distance Education-CBCS-(2019-20 Academic Year Onwards)</u>

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.

UNIT-III CURRICULAM

Structure

1.1 Introduction

Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes. In this unit we mainly focused on the sequence analysis, sequency assembly, pharwise sequence comparison and genome sequence annotation. In this unit will be helpful to analyse the sequence and their relationship. The sequence assembly and sequencing will be used to gain the knowledge about the various methods for constructing whole genome sequence. Hence, this unit elaborates the genomics and their methods.

1.2 Objectives

After going through the unit you will able to;

- Understand the sequence analysis technique
- To identify the whole genome sequencing methods
- Working procedure on pair-wise alignment
- Identify the importance and key role of genome annaotation

1.3 Genomics

Genomics is a forum for describing the development of genome-scale technologies and their application to all areas of biological investigation. That has evolved with the field that carries its name, Genomics focuses on the development and application of cutting-edge methods, addressing fundamental questions with potential interest to a wide audience.

1.3.1 The scope of Genomics

Genome Biology covers all areas of biology and biomedicine studied from a genomic and post-genomic perspective.

It should be used in sequence analysis; bioinformatics; insights into molecular, cellular and organismal biology; functional genomics; epigenomics; population genomics; proteomics; comparative biology and evolution; systems and network biology; genome editing and engineering; genomics of disease; and clinical genomics. • Genomics including genome projects, genome sequencing, and genomic technologies and novel strategies.

• Functional genomics including transcriptional profiling, mRNA analysis, microRNA analysis, and analysis of noncoding and other RNAs using established and newly-emerging technologies (such as digital gene expression).

• Evolutionary and comparative genomics, including phylogenomics.

• Genomic technology and methodology development, with a focus on new and exciting applications with potential for significant impact in the field and emerging technologies.

• Computational biology, bioinformatics and biostatistics, including integrative methods, network biology, and the development of novel tools and techniques.

• Modern genetics on a genomic scale, including complex gene studies, population genomics, association studies, structural variation, and gene-environment interactions.

• Epigenomics, including DNA methylation, histone modification, chromatin structure, imprinting, and chromatin remodeling.

• Genomic regulatory analysis, including DNA elements, locus control regions,

insulators, enhancers, silencers, and mechanisms of gene regulation.

• Genomic approaches to understanding the mechanism of disease pathogenesis and its relationship to genetic factors, including meta-genomic and the mode and tempo of gene and genome sequence evolution.

• Medical Genomics, Personal Genomics, and other applications to human health.

• Application of Genomic techniques in model organisms that may be of interest to a wide audience.

• RNA-seq experiments without three or more biological replicates will not be accepted for differential expression-based articles.

1.4 Sequence analysis

From gene sequences to the proteins they encode to the complicated biological networks they are involved in, computational methods are available to help analyze data and formulate hypotheses. So they have focused on commonly used software packages, to attempt to encompass every detail of every program out there.

The first tools they describe are those that analyze protein and DNA sequence data. Sequence data is the most abundant type of biological data available electronically. While other databases may eventually rival them in size, the importance of sequence databases to biology remains central. Pairwise sequence comparison is the most essential technique in computational biology. It allows doing everything from sequence-based database searching, to building evolutionary trees and identifying

characteristic features of protein families, to creating homology models. But it's also the key to larger projects, limited only by our imagination—comparing genomes, exploring the sequence determinants of protein structure, connecting expression data to genomic information, and much more.

The types of analysis that can do with sequence data are:

- Knowledge-based single sequence analysis for sequence characteristics
- Pairwise sequence comparison and sequence-based searching
- Multiple sequence alignment
- Sequence motif discovery in multiple alignments
- Phylogenetic inference

Check your progress-1

Note: a. writes your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. What is Sequence analysis?

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

1.5 Genome Sequencing

1.5.1 History of sequencing

British biochemist named Frederick Sanger, laid the foundation for sequencing proteins. In 1955, Sanger had completed the sequence of all the amino acids in insulin. His work provided evidence that proteins consisted of chemical entities with a specific pattern, rather than a mixture of substances.

Later, a method named as Sanger Sequencing was developed by Frederick Sanger and his colleagues in 1977, where DNA could be sequenced by generating fragments. It was the most widely used sequencing method for approximately 40 years.

1.5.2 Whole Genome Sequencing and Sequence Assembly

A DNA sequencing reaction produces a sequence that is several hundred bases long. Gene sequences are typically thousands of bases long. The largest known gene is the one associated with Duchenne muscular dystrophy. It is approximately 2.4 million bases in length.

The basic procedure in sequencing has been to isolate genomic DNA or RNA (reverse transcribed into cDNA), and clone it into vectors (eg. plasmids, BACs) capable of stable propagation in suitable host cells such as Escherichia coli. Several cloning systems with insert sizes varying from hundreds of base pairs to megabases have been developed. The ideal clone library for genomic sequencing has the following features.

- The clones are highly redundant, covering the entire genome many times (typically 6–10).
- The clone coverage is random and not biased towards or against specific regions of the genome.

- The clones are stable, not subject to recombination or reorganization during the propagation process.

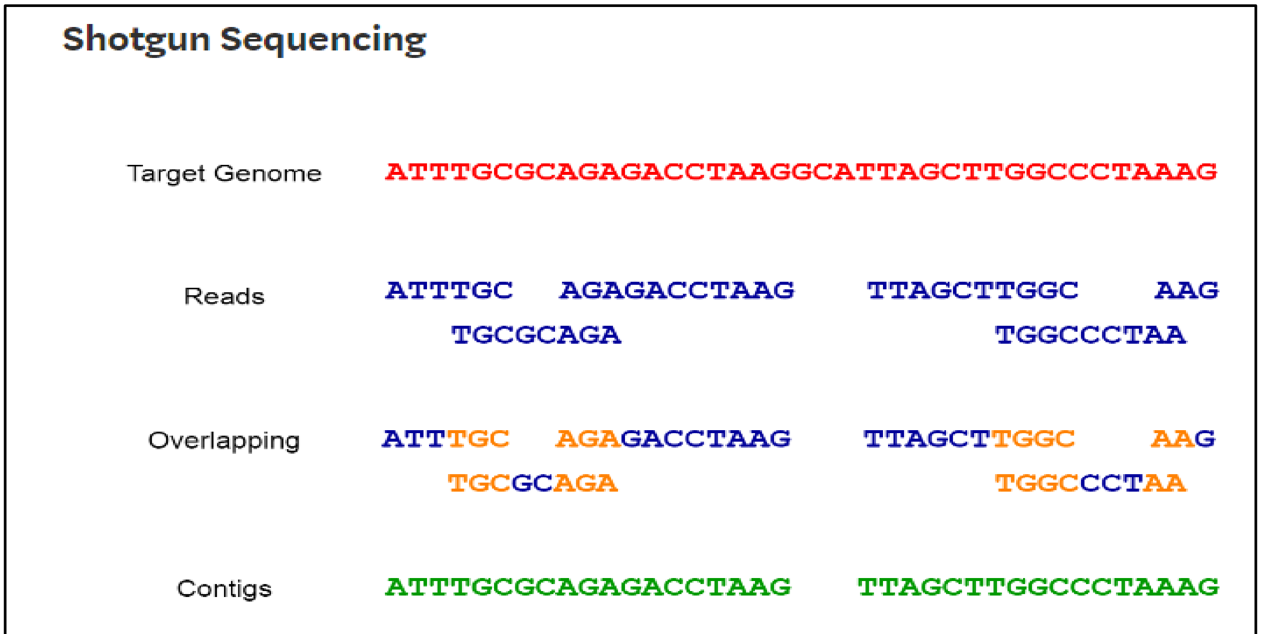
It should be noted that one of the major improvements of the new massively parallel sequencing technologies is that they do not rely on vector cloning prior to sequencing, and the concerns listed here are therefore not directly applicable to those technologies. After propagation, the clones are selected and the sequencing is performed. An essential feature in sequencing is the attachment of quality values to the raw sequences. The quality values indicate the likelihood of each base call being correct. In the assembly stage the quality values will help to distinguish true DNA polymorphisms from sequencing errors and match end sequences of low quality. In genomic shotgun sequencing, which typically uses a single individual DNA source, sequences sharing less than 98% identity are usually assumed to come from different regions of a genome (including different repetitive elements). In contrast, EST data is usually derived from a variety of sources representing the spectrum of polymorphisms in the original samples. These will usually include a number of erroneous polymorphism which are caused by sequencing errors inherent in single pass sequencing, a relatively high rate of insertions and deletions, contamination by vector and linker sequences and the non-random distribution of sequence start sites in oligo(dT)-primed libraries. Therefore, the degree of identity in overlapping sequences from the same gene will often be lower in EST projects than in genomic sequencing projects. In addition, the patterns of overlapping sequences caused by alternative splice forms are different from those observed in a genomic shotgun project.

1.5.3 Shotgun sequencing

Two approaches for genome shotgun sequencing can be distinguished: whole-genome shotgun (WGS) sequencing and hierarchical shotgun sequencing.

Shotgun sequencing

Sequencing using the whole genome shotgun approach basically means that the genome is randomly broken into pieces and cloned into a sequencing vector. The inserts are subsequently processed to generate sequences of bases (referred to as reads). During the mid 1990s several groups recognized that sequence information from both ends of relatively long inserts dramatically improves the efficiency of sequence assembly. In contrast to single sequence reads from one end of the shotgun clones pairs of sequence reads from both ends have known spacing and orientation. Exact knowledge of the length of the insert is not required to utilize the advantages of end sequencing in assembly, but good estimates of clone length will aid the assembly immensely.



Shotgun Sequencing

Hierarchical shotgun sequencing

The 'Hierarchical shotgun sequencing' (also referred to as 'map-based', 'BAC-based' or 'clone-by-clone') approach involves generating and organizing a set of large insert clones (typically 100–200 kb each) covering the genome (a "minimal tiling path"), followed by separate shotgun sequencing on each clone. It is possible to establish a tiling path of overlapping BAC-clones using only BAC fingerprinting technologies. However, knowledge of unique genome markers (*eg.* ESTs or sequence-tagged sites (STS)) and their location in the genome map is of great help for organizing the BAC clones in the correct order. In hierarchical shotgun sequencing the sequence information is local; therefore the risk of long-range and short-range misassembly is reduced.

Mixed strategy sequencing

A strategy that can be used on large complex genomes is the ’mixed strategy sequencing’. The technique utilizes both hierarchical and whole-genome shotgun. The method combines a light (x1) BAC clone coverage of the genome, with whole genome shotgun sequencing. The BAC clones act as a basic framework for WGS sequence assembly. The method was successfully applied to rat genome.

Reduced Representation Sequencing

A variant of WGS is “reduced representation sequencing” (RRS), where one selectively chooses subsets of the genome to avoid sequencing the (often much) larger regions that are not of interest. In, SNPs were discovered by mixing DNA from many individuals, preparing a library of appropriately sized restriction fragments, and randomly sequencing 5 clones. Here, the choice of the restriction fragments effectively selects only a small subset of the human genome. Several approaches to RRS have been employed for plant genomes: Methyl-filtration (MF) sequences uses the endogenous restriction-modification system of *E. coli* to eliminate methylated DNA inserts, the RescueMu (RM) approach focuses on the gene-rich regions which are rich in mutator transposons, and High-Cot filtration avoids repetitive and low copy sequences due to differences in the relative rates of DNA re-association. Most of the Maize and Sorghum genomes have been sequenced using MF. Many of the applications of the new high throughput sequencing platforms are based on various RSS strategies. This includes electrophoretic size separation to enrich for small RNA molecules; reduced representation bisulphite sequencing for genome wide methylation analysis; flow sorting of derivative translocation chromosomes for breakpointmapping; enrichment of DNA-fragments bound to specific proteins by chromatin immunoprecipitation of fixed, sheared DNA, for identification of transcription factor binding

sites (CHiP-Seq); enrichment of Specific parts of the genome by multiplex PCR-amplification or by hybridization to custom made arrays , *eg.* For SNP discovery and in situ exon capture.

EST sequencing

EST (expressed sequence tag): A unique stretch of DNA within a coding region of a gene that is useful for identifying full-length genes and serves as a landmark for mapping. An EST is a sequence tagged site (STS) derived from cDNA.

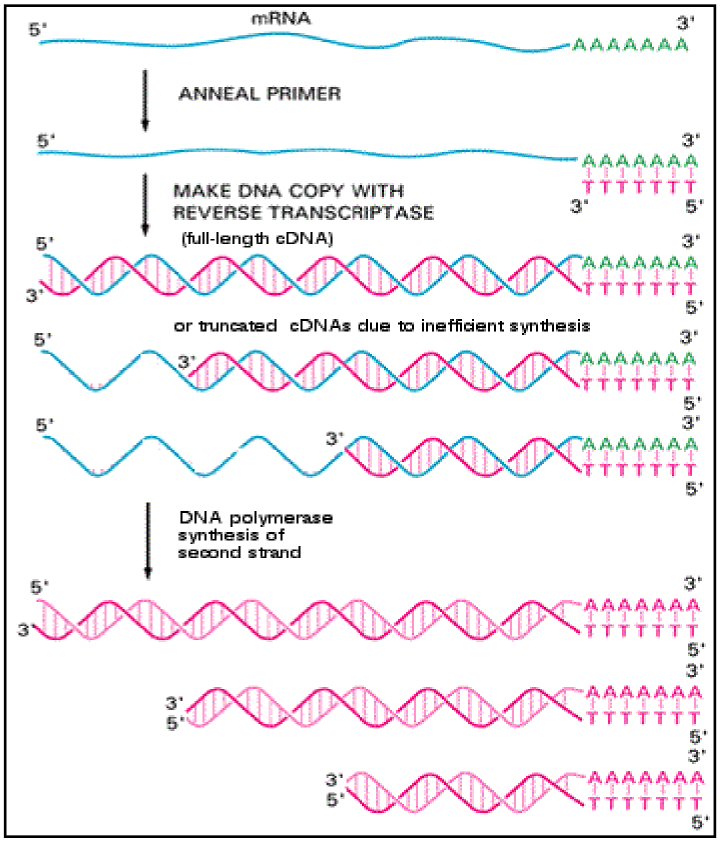
An STS is a short segment of DNA which occurs but once in the genome and whose location and base sequence are known. STSs are detectable by the polymerase chain reaction (PCR), are helpful in localizing and orienting mapping and sequence data, and serve as landmarks in the physical map of the genome.

Expressed Sequence Tags (ESTs) are sequences representing genes which can originate from specific tissues. In EST-sequencing a single automated sequencing from one or both ends of cDNA-inserts is performed. This singlepass approach is the major reason EST-sequencing is cost effective. In most cases EST sequencing projects are aimed at establishing partial sequences of transcribed genes rather than full length cDNA sequences. However, this approach features some special challenges such as common sequence motifs, alternative transcripts and paralogous genes are challenges that potentially impact the assembly quality.

Single sequence reactions from cDNA libraries of specific tissues and culture conditions

- readout of mRNA sequences present in cell

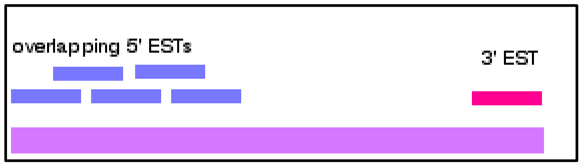
cDNAs are synthesized from mRNA mixture present in tissue of interest using reverse transcriptase.



Once cDNAs are synthesized, they are cloned into a vector for sequencing. Usually, they are cloned in a known orientation, so that sequence can be derived from the 3' end of the message or 5' end. Since full-length cDNAs are difficult to synthesize, different cDNA copies of the same mRNA may have different sequences at their 5' ends.

- so multiple cDNAs of a particular message can give overlapping reads of mRNA sequence
- can assemble into larger "contig" of mRNA sequence by clustering (similar to genomic sequence assembly)

Clustering



Some ESTs may differ in sequence, due to variations in mRNA splicing. Sequences are assembled onto known full-length cDNAs from genbank, if known. Assembly is called a UniGene - unique gene from which ESTs are derived.

Once UniGenes determined, the EST can be used as a measure of gene expression within a tissue:

- Highly expressed genes will be represented by numerous different ESTs
- Low expression genes will only rarely show up in random library of ESTs

But this means a few genes can be represented by most of the ESTs,

- need to sequence many more ESTs to discovered novel sequences

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Comment on different types of Sequencing methods

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

1.6 Sequence Assembly

Genome sequencing is a discipline that has undergone tremendous development in the past. Today 567 bacterial genomes with up to 10.5 million base pairs (Plesiocystis pacifica SIR-1) have been sequenced.

Sequence assembly is the only way to efficiently assemble sequenced fragments of DNA. However, a sufficient amount of high quality sequences are required. The assembly programs should be able to handle large data sets effectively and avoid misassemblies in the presence of large repetitive or duplicated regions and redundant sequences. To accomplish this, effective algorithms to handle large input data sets with the use of minimal computer time and memory are needed. One of the primary difficulties in computational genome assembly is to develop an algorithmic approach capable of detecting stretches of repetitive DNA without causing misassemblies. Repetitive sequences complicate assembly as different pieces of sequence can share the same repeat sequence originating from different genomic locations. Since the pieces are put together by searching for matching overlapping nucleotides, repeats can be put together erroneously. Typically, for shotgun data, repetitive sequences are revealed by clusters containing more overlapping reads than would be expected by chance.

In EST datasets the main difficulty is to develop an algorithmic approach that, in addition to efficient assembly, can handle highly expressed genes, paralogous genes, alternative spliceforms and chimerism in the dataset. The theoretical background for genome assembly lies in computer science, and an insight into the mathematical and theoretical background can be found in and references therein. Although pyrosequencing with a whole-genome shotgun approach has been successfully applied to bacterial genomes, the construction of high-quality assemblies with high-

throughput sequencing data is still a non-trivial problem even for short genomes. At present, no approach has been proposed to directly assemble large animal or plant genomes directly from short sequences obtained using HTS. As described below the Short Read Assembly Protocol (SHRAP), however, comprises a protocol for high-throughput short read sequencing that differs in two respects from classical hierarchical sequencing approaches. This protocol however, expects read lengths much longer (200 nucleotides) than those produced by SOLiD or Solexa. The assembly methodology is based on the Euler engine introduced in 2004.

The Euler-SR assembler, specifically designed to assemble short reads, uses an updated version of the Euler engine to reduce memory requirements. The results for real Solexa reads, however, were less convincing due to the poorly understood error model and highly variable error rates across different machines and run times.

1.6.1 Basic principles of Assembly

For the majority of traditional assembly programs the basic scheme is the same, namely the overlap-layout-consensus approach.

Essentially it consists of the following steps:

• Sequence and quality data are read and the read are cleaned.

• Overlaps are detected between reads. False overlaps, duplicate reads, chimeric reads and reads with self-matches (including repetitive sequences) are also identified and left out for further treatment.

• A multiple sequence alignment of the reads is performed, and a consensus sequence is constructed for each contig layout (often along with a computed quality value for each base).

• Possible sites of misassembly are identified by combining manual inspection with quality value validation. Prior to the assembly, the electropherogram(for Sanger sequencing, images for massively parallel sequencers) for a given sequence is interpreted as a sequence of bases (a read) with associated quality values, these values reflect the log-odds score of the bases being correct. The basecaller PHRED is often used, however alternatives exist, *eg.* The CATS basecaller.

The reads can then be screened for any contaminant DNA such as *Escherichia coli*, cloning or sequencing vector. Low quality regions can be identified and removed. Base quality values can be used in computation of significant overlaps and in construction of the multiple alignments.

For high-throughput sequencing data, the basic proposition for SHRAP is to sample clones from the genome at high coverage, while sequencing reads from these clones at low coverage. SHRAP starts off with assembling the reads greedily to small local assemblies and subsequently to contigs on each clone. It proceeds by ordering the clones in a "clone graph", and constructing "clone contigs", which are then assembled independently. Computer simulations of the procedure show that the approach can reach a quality comparable to the current assemblies of single human chromosomes and fruit fly genomes using reads of 200nt with an error rate of not more than 1%. These are constraints that are too strict for short (Solexa or SOLiD) reads (≈ 40*bp*) and because of higher error rates challenging for real 454 reads. Furthermore, for mammalian genomes the use of a hierarchical sequencing strategy might be somewhat cumbersome. However, the use of templates might bail Solexa and SOLiD users out: In a recent study, de novo assemblies of chloroplast genomes (≈ 120 kb) were improved by aligning preassembled contigs to reference genomes. After de-Bruijn graph assembly of reads, small contigs were aligned to closely related chloroplast genomes. Between 67% and 98% of the contigs could be aligned to such templates. If alignment failed, sequences were scanned for

similarity using BLASTN. The authors reported that successful BLASTN matches typically contained > 100 bp insertions relative to the reference genome. In the end, however, their assemblies were estimated to be 88–94% complete. The successful de novo assembly of Chloroplasts genomes with 454 reads has been shown earlier.

SOLiD, Solexa technologies allow convenient generation of mate-pair/paired-end sequences, *ie.* The ability to sequence both ends of each DNA fragment. However, in an assembly using a hybrid dataset of real 454 reads and simulated mate-pair data, about 96% of the mate-pairs did not contribute additional information and hence did not improve the assembly. Likewise, in a hybrid dataset of 454 reads and Sanger reads the vast majority of long sequences did not improve the assembly substantially, measured by N50 contig size. Hence, the authors concluded that hybrid protocols should be reviewed critically. Despite those simulation results, the latter method has already been shown to work quite well in practice, and one area wheremate-pair/paired-end sequencing should improve the analysis dramatically is for the detection of breakpoints related to structural rearrangements, *eg.* Deletions, duplications, inversions and translocations

1.7 Pairwise Sequence Comparison

Pair-wise alignment produces an optimal alignment using a scoring matrix for the given two sequences

Description

Pairwise alignment is an arrangement in columns of 2 sequences, highlighting their similarity.

Input

Protein sequence is given as input.

G A A T T C

G A T T A

Procedure

1. Go to http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html
2. Paste the protein sequence and click compute parameters.
3. The result page is retrieved.

Output

```
########################################
# Program: needle
# Rundate: Fri 13 May 2016 12:03:53
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20160513-120351-0311-34620455-oy.asequence
#    -bsequence emboss_needle-I20160513-120351-0311-34620455-oy.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 6
# Identity:       3/6 (50.0%)
# Similarity:     3/6 (50.0%)
# Gaps:           1/6 (16.7%)
# Score: 7.0
#
#
#=======================================

EMBOSS_001         1 GAATTC      6
                      .|||.
EMBOSS_001         1 -GATTA      5
```

Output for Pair wise Alignment

1.8 Annotating and analyzing genome sequences

DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary

Genome annotation consists of three main steps:

1. Identifying portions of the genome that do not code for proteins
2. Identifying elements on the genome, a process called gene prediction
3. Attaching biological information to these elements

Automatic annotation tools attempt to perform these steps via computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

A simple method of gene annotation relies on homology based search tools, like BLAST, to search for homologous genes in specific databases; the resulting information is then used to annotate genes and genomes. However, as information is added to the annotation platform, manual annotators become capable of deconvoluting discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g. Ensembl) rely on curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- gene structure
- coding regions
- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. Proteogenomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.

A variety of software tools have been developed to permit scientists to view and share genome annotations; for example, MAKER.

Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals (The 100,000 Genomes Project,UK)and several model organisms are largely complete. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".

Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

Encyclopedia of DNA elements (ENCODE)

Entrez Gene

Ensembl

GENCODE

Gene Ontology Consortium

GeneRIF

RefSeq

Uniprot

Vertebrate and Genome Annotation Project (Vega)

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Note down the two types of annotation

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

1.9 LET US SUM UP

In this unit, you have learnt about the genome sequencing assembly, analayis and annotation. This knowledge would make understand how the sequence is constructed and various analyzing techniques. The sequence assembly and annotation would pave the key attention to the student for the structural and functional aspect of student. The sequencing techniques are learnt.

2.0 ANSWERS TO CHECK YOUR PROGRESS

1**.** Sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

2. Shotgun sequencing, Hierarchical shotgun sequencing, Mixed strategy sequencing, Reduced Representation Sequencing and EST sequencing

3. Structural and functional Annotation

<u>Distance Education-CBCS-(2019-20 Academic Year Onwards)</u>

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.

UNIT-IV CURRICULAM

Structure

1.1 Introduction

1.2 Objectives

1.3 Genbank sequence queries against biological databases

1.4 Basic Local Alignment Search Tool (BLAST)

1.5 FASTA

1.6 Multifunctional Tools for Sequence Analysis

- 1.6.1 NCBI SEALS
- 1.6.2 The Biology Workbench
- 1.6.3 DoubleTwist

1.7 Let Us Sum Up

1.8 Answers to check your progress

1.1 Introduction

Sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. In this unit we will discuss about the sequence related databases, BLAST and FASTA analysis, and multifunctional tool analysis. This unit is mainly helpful to predict the function of unkown sequence. The genbank database information is used to gain the knowledge about various tools and sequence database annotation and etc..

1.2 Objective

- Sequence submission and retrieval form Genbank will be learnt.
- To identify the methods and procedure of BLAST and FASTA tools
- Know about the multifunctional tools for sequence analysis

1.3 Genbank sequence queries against biological databases

Genbank, was built and maintained by the National Center for Biotechnology Information (NCBI). It is part of the International Nucleotide Sequence Database in collaboration with the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences, acquire initially through submissions from individual laboratories and batch submissions from large-scale sequencing projects. NCBI's retrieval system helps to access GenBank and Entrez used to combine information from the major DNA and protein sequence databases along with mapping, taxonomy genome, protein structure and domain information, and the literature information through PubMed (McEntyre and

Lipman, 2001). The NCBI has integrated server called BLAST (Basic Local Alignment Searching Algorithm) that identifies similar sequences for the query sequence. Up-to-now nearly 60 million sequences are available and it's growing aggressively. Genbank also provides the FTP access to the user wher they can retrieve the large set of data for their analysis from ftp://ftp.ncbi.nih.gov. Each sequence in GenBank is annotated with literature reference, size, sequence features, organism, and protein translations. The GenBank sequences are arranged under several divisions *viz.* EST, ENV, Plant sequences (PLN), Genome survey sequences (GSS), rodents (ROD), Sequence tagged site (STS) *etc*.,

GenBank entry includes a brief description of the sequence, the scientific name, and taxonomy of the source organism, bibliographic references, and a table of features (http://www.ncbi.nlm.nih.gov/collab/FT/index.html). Table 1. lists different sequence divisions in Genbank.

Table 1. Different division under which sequences are made available in GenBank

Code	**Description**
PRI	primate sequences
ROD	rodent sequences
MAM	other mammalian sequences
VRT	other vertebrate sequences
INV	invertebrate sequences
PLN	plant, fungal, and algal sequences
BCT	bacterial sequences
VRL	viral sequences
PHG	bacteriophage sequences
SYN	synthetic sequences
UNA	unannotated sequences
EST	EST sequences (expressed sequence tags)

PAT	patent sequences
STS	STS sequences (sequence tagged sites)
GSS	GSS sequences (genome survey sequences)
HTG	HTGS sequences (high throughput genomic sequences)
HTC	HTC sequences (high throughput cDNA sequences)
ENV	Environmental sampling sequences

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. What is Genbank?

……………………………………………………………………………………………

1.4 Basic Local Alignment Search Tool (BLAST)

NIH has developed the BLAST tool. It uses statistical methods to estimate hits for their significance. This program recognizes sequences in the database that share common words of a pre-set size (k-tuple) and these matching words are expanded to identify un-gapped common segments between two sequences. A brief note on the alignment statistics and implementation of the BLAST search method is available at National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). There are many variants such as blastx or tblastn can efficiently used to search DNA sequence or translated sequences. BLAST search collects closely related sequences with high sequence similarity. Divergent sequences and an initial search produces no or very few hits can be efficiently managed by Higher order PAM

(PAM250) or low order BLOSUM (BLOSUM40) matrices. The weaker similarities can be matched using PSIBLAST by an iterative position specific scoring matrix. In the first iteration PSIBLAST also function as BLAST. However, once sequences with specified E-value cutoff are identified a position specific matrix (PSSM) can be calculated in order to identify other related sequences. A more elegant note on the implementation algorithm is described at the website http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html.

1.5 FASTA

The shared regions in two sequences and score sequences in a database for homology can rapidly be identified by FASTA program (Pearson, 1998). The final output consists of a rank ordered list of sequences and alignment between sequences. The high similarity regions among the sequences are identified by segments with high frequency of conserved letters (ktup). A general value set for protein sequence search is 2 and it will look for pairs of conserved alphabets. Needleman-Wunch-Sellers algorithm is used to calculate the optimized score. The FASTA can also be used for DNA and protein sequence search. A web interface to FASTA tool that can search sequence database is available at http://fasta.bioch.virginia.edu/. Different databases such as PDB (http://www.rcsb.org) and GENOME DB (http://fasta.genome.jp) also provide FASTA search facility.

Check your progress-2

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Define BLAST and FASTA

1.6 Multifunctional Tools for Sequence Analysis

1.6.1 NCBI SEALS

The NCBI SEALS project aims to develop a Perl-based command-line environment for Systematic Analysis of Lots of Sequences. SEALS is far from a fully automated genome analysis tool, and it isn't intended to be. What SEALS does provide is an enhancement to the command-line environment on Unix systems. It is composed of a large suite of scripts with a variety of useful functions: converting file formats, manipulating BLAST results and FASTA files, database retrieval, piping files into Netscape, and a host of other features that make your data easier to look at without requiring a resource-sucking GUI. SEALS runs on Unix systems and is probably most useful for those who are already Unix aficionados. Before you write a script to process a lot of sequences, check to see if the process you want has been implemented in SEALS.

1.6.2 The Biology Workbench

The San Diego Supercomputing Center offers access to sequence-analysis tools through the Biology Workbench. This resource has been freely available to academic users in one form or another since 1995. Users obtain a login and password at the SDSC site, and work sessions and data can be saved securely on the server.

The Biology Workbench offers keyword and sequence-based searching of nearly 40 major sequence databases and over 25 whole genomes. Both BLAST and FASTA are implemented as search and alignment tools in the Workbench, along with several local and global alignment tools, tools for DNA sequence translation, protein sequence feature analysis, multiple sequence alignment, and phylogenetic tree drawing. The Workbench group has not yet implemented

profile tools, such as MEME, HMMer, or sequence logo tools, although PSI-BLAST is available for sequence searches.

Although its interface can be somewhat cumbersome, involving a lot of window scrolling and button clicking, the Biology Workbench is still the most comprehensive, convenient, and accessible of the web-based toolkits. One of its main benefits is that many sequence file formats are accepted and translated by the software. Users of the Workbench need never worry about file type incompatibility and can move seamlessly from keyword-based database search, to sequence-based search, to multiple alignment, to phylogenetic analysis.

1.6.3 DoubleTwist

Another entry into the sequence analysis portal arena is DoubleTwist at http://doubletwist.com. This site allows you to submit a sequence for comparison to multiple databases using BLAST. It also provides "agents" that monitor databases for new additions that match a submitted sequence and automatically notifies the user. These services, as well as access to the EcoCyc pathways database and to an online biology research protocols database, are free with registration at the site at the time of this writing**.**

Check your progress-3

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. List out the multifunctional tools for sequence analysis

………

……………………………………….

1.7 Let us sum up

GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences which is useful to student for gaining knowledge to every organisms. The relationship between the two sequences is identified through BLAST and FASTA analysis. Student can know about the function of unknown sequence they learn to analyze the sequence.

1.8 Answers to check your progress

1. GenBank is a comprehensive database that restrains more than 1,65,000 named organisms DNA sequences, acquire initially through submissions from individual laboratories and batch submissions from large-scale sequencing projects.
2. BLAST search collects closely related sequences with high sequence similarity and The shared regions in two sequences and score sequences in a database for homology can rapidly be identified by FASTA program.
3. NCBI SEALS, The Biology Workbench and DoubleTwist

<u>Distance Education-CBCS-(2019-20 Academic Year Onwards)</u>

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.

UNIT-V CURRICULAM

Structure

1.1 Introduction

1.2 Objectives

1.3 Multiple-Sequence Alignment

- 1.3.1 ClustalW
- 1.3.2 TCoffee

1.4 Phylogenetic alignment

- 1.4.1 Steps involved in Phylogenetic analysis

1.5 Motifs

1.6 Profile

1.7 Let Us Sum Up

1.8 Answers to check your progress

1.1 Introduction

Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column. In this unit, we are mainly focusing the multiple sequence alignment and phylogenetic alignment. Hence, it is essential to find the ancestor of unknown sequence. Based on this identification we can easily predict the characteristic and function of a sequence.

1.2 Objectives

- To perform multiple sequence alignment with various tools
- To construct phylogenetic tree between the sequences
- Identify the motif and profile.

1.3 Multiple-Sequence Alignment

Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column. The principle of the BLAST or FASTA can identify closely related sequences but it is not provide information about conserved blocks of residues in a protein family. The main application of creating multiple sequence alignment of several protein sequences is to identify the significant conserved regions and in many cases to understand function. Heuristic methods based on progressive-alignment (ClustalW, TCOFFEE), simultaneous alignment of all sequences (MSA), using iterative strategies (Prrp) are the major contributors to produce multiple global sequence alignments (Higgins and Sharp, 1988).

1.3.1 ClustalW

ClustalW is a multiple sequence alignment tool for protein or DNA sequences. The closely related sequences are added and more evolutionarily diverged sequences are added in this algorithm. Two different steps to produce a complete multiple alignment are of (1) Identification of closely related sequences among the input, (2) Ordering these sequences based on pair-wise similarity score, (3) Alignment construction using those sequence pairs with highest similarity score, (4) Addition of sequences in a tree-order manner to create alignments. EMBL website contains a web interface at http://www.ebi.ac.uk/clustalw/.

1.3.2 TCoffee

TCoffee is a multiple alignment tool based on progressive-alignment strategy on a scoring matrix. A global and local pair-wise alignment of all sequences is created as a first step and weights for individual residue pairs are assigned based on the percentage identity of aligned residues. Further, using this library of primary alignments, an extended library of alignments has generated by considering all possible primary alignments. A correct alignment is generated using dynamic programming. A correct multiple alignment should characterize a structural alignment among input sequences. BaliBase database has been used to test the TCoffee accuracy of alignment (Thomopson *et al*., 1999). T-Coffee program can be accessed at http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi. Both TCoffee and Clustal methods are included in the BioPerl package (http://www.bioperl.org).

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Define Multiple Sequence alignment

………………………………………………………………………………………………

1.4 Phylogenetic alignment

Phylogenetics is the study of evolutionary relationships. The evolutionary history deduced from phylogenetic analysis is usually represented as branching, tree like diagrams that represent an estimated pedigree of the inherited relationships among molecules (“gene trees”), organisms. Phylogenetics is sometimes called as cladistics. Sequences are outshined morphological and other organismal characters as the most popular form of data for phylogenetic or cladistic analysis. Numerous phylogenetic algorithms, procedures, and computer programs have been invented and their reliability and practicality are mainly on the structure and size of the data.

1.4.1 Steps involved in Phylogenetic analysis

1. Alignment
2. Determining the substitution model
3. Tree building
4. Tree evaluation

Alignment (Building the data model)

Multiple sequence alignments is the major part of phylogenetic sequence alignments. The alignment based positions are referred as “sites”. These sites are equivalent to “characters”.

Determining the substitution model

The substitution model is also very important as like as alignment and tree building. This model manipulates the alignment and tree building. The model of substitution can be generated either from particular bases or from the relative rate of overall substitution among different sites in sequences.

Tree building methods

Tree building methods can be arranged into character based Vs distance based methods. Character based method can be derived as the optimization of the distribution of the actual data patterns for each character. The distance based methods compute the pairwise distances according to some measurement to derive trees.

Tree evaluation

Several procedures are available that evaluate the phylogenetic signal in the data and the robustness of trees. There are tests of data signal versus randomized data (skewness and permutation tests) is the most popular method to evaluate the constructed phylogentic tree. The likelihood ratio test provides a means of evaluating both the substitution model and the tree.

1.5 Motifs

A related set of nucleotides or residues may possess conserved patterns may have specific functions. These conserved patterns are known as motifs. The transcription factors may bind to specific nucleotide motifs in DNA sequences. The initial step of this process is to collect a set of sequences with common function. These sequences may belong to the same protein family or may be diverse with the common function. PROSITE is a widely used database of patterns or motifs with details of function and sequence contain these patterns. Currently it contains 1309 patterns. Commonly occurring patterns are expressed with IUPAC single letter amino acid code. Square brackets represent positions where more than one amino acid. Excluding a set of amino acids at a particular position represented in a curly braces. Minimum and maximum repetition of residues is represented in numbers inside a bracket. PROSITE patterns can be downloaded in http://www.expasy.ch/prosite/ from ExPasy website. Patterns can be identified for a given sequence in PROSITE database using search engine "ScanProsite". Two different methods viz.

physical-chemical properties and scoring the occurrence of amino acids with higher rate of substitution were used to score the conservation of residues in a particular position.

1.6 Profile

Profile is a quantitative or qualitative method of describing a motif. A profile can be expressed in its most rudimentary form as a list of the amino acids occurring at each position in the motif. Early profile methods used simple profiles of this sort; however, modern profile methods usually weight amino acids according to their probability of being observed at each position.

Check your progress-2

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Explain Phylogenetic analaysis

………

…………………………………………

1.7 Let us sum up

In this unit, you have learnt about the multiple sequence and phylogentic alignment methods and procedures. This knowledge make you understand what multiple sequence alignment is and how it can be analyzed in various tools. The concept such phylogentic alignment and tree construction would have made you to generating phylogenetic tree. This content might play very important role in phylogenetic tree construction.

1.8 Answers to check your progress

1. Multiple-Sequence alignment produces an optimal alignment using a scoring matrix for the given two sequences whereas, multiple sequence alignment optimally aligns several related sequences and evolutionarily constrained residues are aligned under the same column.
2. Phylogenetics is the study of evolutionary relationships. The evolutionary history deduced from phylogenetic analysis is usually represented as branching, tree like diagrams that represent an estimated pedigree of the inherited relationships among molecules (“gene trees”), organisms.

Distance Education-CBCS-(2019-20 Academic Year Onwards)

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours | Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.

UNIT-VI CURRICULAM

Structure

1.1 Introduction

Proteomics is the analysis of the entire protein complement of a cell, tissue, or organism under a specific, defined set of conditions. In this unit, we cover the protein structure determination and prediction. The determination of protein structure denotes how the protein structures were solved by experimentally. The PDB and SWISSPROT database have the information about protein structure and function. Homology modeling is used predict the protein structure by bioinformatics approaches. This unit gives a vast attention about the protein structures to the student. This unit might be helpful to the learner to understand the protein structure and function.

1.2 Objectives

1. Retrieve the protein structure from PDB and SWISSPROT database
2. To understand the structure determination of protein
3. To predict the secondary and 3D structure of protein
4. Analyze the biochemical pathway of rel;ated protein

1.3 Protein Data Bank

The Protein Data Bank is the single worldwide repository of large molecules of proteins and nucleic acids. It is available at http://www.pdb.org/. The PDB contains 1,25,180 protein structures and 2,895 nucleic acid structures. This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that help students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

1.4 SWISS-PROT

SWISS-PROT is a protein knowledge based database that contains amino acid sequence, protein name, taxonomic data, and citation information. It also offers cross-references to external

data collections such as the fundamental DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence database, 2D and 3D protein structure databases, post translational modification (PTM) databases, various protein domain and family characterization databases, species-specific data collections, disease databases and variant databases. SWISS PROT is steadily being improved by the addition of a number of features from OMIM GeneLynx, GeneCards, Genew as well as to many gene-specific mutation databases. SWISS-PROT and TrEMBL can be obtained by anonymous FTP from the ExPASy server ftp://expasy.org and EBI server ftp://ebi.ac.uk/pub/. Further, user can obtain weekly updates and complete data sets in various formats from http://www.expasy.org/sprot/download.html.

Check your progress-1

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Define PDB and Swiss prot

………………………………………………………………………………………………

1.5 Biochemical Pathway Databases

1.5.1 WIT and KEGG

WIT is a metabolic pathway reconstruction resource; that is, the curators of WIT are attempting to reconstruct complete metabolic pathway models for organisms whose genomes have been completely sequenced. WIT currently contains metabolic models for 39 organisms. The WIT

models include far more than just metabolism and bioenergetics; they range from transcription and translation pathways to transmembrane transport to signal transduction.

The General Search function allows you to search all of WIT, or subsets of organisms. This type of search is based on keywords, using Unix-style regular expressions to find matches. There is also a similarity search function that allows you to search all the open reading frames (ORFs) of a selected organism for sequence pattern matches, using either BLAST or FASTA. Pathway queries require you to specify the names of metabolites and/or specific EC enzyme codes. Enzyme queries allow you to specify an enzyme name or EC code, along with location information such as tissue specificity, cellular compartment specificity, or organelle specificity. In all except the regular-expression searches, the keywords are drawn from standardized metabolic vocabularies. WIT searches require some prior knowledge of these vocabularies when you submit the query. WIT was primarily designed as a tool to aid its developers in producing metabolic reconstructions, and documentation of the vocabularies used may not always be sufficient for the novice user. WIT is relatively text-heavy, although at the highest level of detail, metabolic pathway diagrams can be displayed.

Another web-based metabolic reconstruction resource is KEGG, which provides its metabolic overviews as map illustrations, rather than text-only, and can be easier to use for the visually-oriented user. KEGG also provides listings of EC numbers and their corresponding enzymes broken down by level, and many helpful links to sites describing enzyme and ligand nomenclature in detail. The LIGAND database, associated with KEGG, is a useful resource for identifying small molecules involved in biochemical pathways. Like WIT, KEGG is searchable by sequence homology, keyword, and chemical entity; you can also input the LIGAND ID codes of two small molecules and find all of the possible metabolic pathways connecting them.

1.5.2 PathDB

PathDB is another type of metabolic pathway database. While it contains roughly the same information as KEGG and WIT—identities of compounds and metabolic proteins, and information about the steps that connect these entities—it handles information in a far more flexible way than the other metabolic databases. Instead of limiting searches to arbitrary metabolic pathways and describing pathways with preconceived images, PathDB allows you to find any set of connected reactions that link point A to point B, or compound A to compound B.

PathDB contains, in addition to the usual search tools, a pathway visualization interface that allows you to inspect any selected pathway and display different representations of the pathway. The PathDB developers plan to incorporate metabolic simulation into the user interface as well, although those features aren't available at the time of this writing.

Check your progress-2

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. List out the Biochemical Pathway Databases

……………………………………………………………………………………………

1.6 Predicting Protein Structure and function from sequence

The amino acid sequence of a protein contains interesting information in and of itself. A protein sequence can be compared and contrasted with the sequences of other proteins to establish its relationship, if any, to known protein families, and to provide information about the evolution of

biochemical function. However, for the purpose of understanding protein function, the 3D structure of a protein is far more useful than its sequence.

The key property of proteins that allows them to perform a variety of biochemical functions is the sequence of amino acids in the protein chain, which somehow uniquely determines the 3D structure of the protein. Given 20 amino acid possibilities, there are a vast number of ways they can be combined to make up even a short protein sequence, which means that given time, organisms can evolve proteins that carry out practically any imaginable purpose.

Each time a particular protein chain is synthesized in the cell, it folds together so that each of the critical chemical groups for that particular protein's function is brought into a precise geometric arrangement. The fold assumed by a protein sequence doesn't vary. Every occurrence of that particular protein folds into exactly the same structure.

Despite this consistency on the part of proteins, no one has figured out how to accurately predict the 3D structure that a protein will fold into based on its sequence alone. Patterns are clearly present in the amino acid sequences of proteins, but those patterns are degenerate ; that is, more than one sequence can specify a particular type of fold. While there are thousands upon thousands of ways amino acids can combine to form a sequence of a particular length, the number of unique ways that a protein structure can organize itself seems to be much smaller. Only a few hundred unique protein folds have been observed in the Protein Data Bank. Proteins with almost completely nonhomologous sequences nonetheless fold into structures that are similar. And so, prediction of structure from sequence is a difficult problem.

1.7 Determination of structure

The protein structure was determined by X-ray Crystallography and NMR Spectroscopy methods

1.7.1 X-ray Crystallography

Linus Pauling and Robert Corey began to use x-ray crystallography to study the atomic structures of amino acids and peptides. In experiments that began in the early 1950s, John Kendrew determined the structure of a protein called myoglobin, and Max Perutz determined the structure of a similar protein called hemoglobin. Both proteins are oxygen transporters, easily isolated in large quantities from blood and readily crystallized. Obtaining x-ray data of reasonably high quality and analyzing it without the aid of modern computers took several years. The structures of hemoglobin and myoglobin were found to be composed of high-density rods of the dimensions expected for Pauling's proposed alpha helix. Two years later, a much higher-quality crystallographic data set allowed the positions of 1200 of myoglobin's 1260 atoms to be determined exactly. The experiments of Kendrew and Perutz paved the way for x-ray crystallographic analysis of other proteins. In x-ray crystallography, a crystal of a substance is placed in an x-ray beam. X-rays are reflected by the electron clouds surrounding the atoms in the crystal. In a protein crystal, individual protein molecules are arranged in a regular lattice, so x-rays are reflected by the crystal in regular patterns. The x-ray reflections scattered from a protein crystal can be analyzed to produce an electron density map of the protein (see Figure 10-1: images are courtesy of the Holden Group, University of Wisconsin, Madison, and Bruker Analytical X Ray Systems). Protein atomic coordinates are produced by modeling the best possible way for the atoms making up the known sequence of the protein to fit into this electron density. The fitting process isn't unambiguous; there are many incrementally different ways in which an atomic structure can be fit into an electron density map, and not all of them are chemically correct. A protein structure isn't an exact representation of the positions of atoms in

the crystal; it's simply the model that best fits both the electron density map of the protein and the stereochemical constraints that govern protein structures.

In order to determine a protein structure by x-ray crystallography, an extremely pure protein sample is needed. The protein sample has to form crystals that are relatively large (about 0.5 mm) and singular, without flaws. Producing large samples of pure protein has become easier with recombinant DNA technology. However, crystallizing proteins is still somewhat of an art form. There is no generic recipe for crystallization conditions (e.g., the salt content and pH of the protein solution) that cause proteins to crystallize rapidly, and even when the right solution conditions are found, it may take months for crystals of suitable size to form.

Many protein structures aren't amenable to crystallization at all. For instance, proteins that do their work in the cell membrane usually aren't soluble in water and tend to aggregate in solution, so it's difficult to solve the structures of membrane proteins by x-ray crystallography. Integral membrane proteins account for about 30% of the protein complement (proteome) of living things, and yet less than a dozen proteins of this type have been crystallized in a pure enough form for their structures to be solved at atomic resolution.

1.7.2 NMR Spectroscopy

An increasing number of protein structures are being solved by nuclear magnetic resonance (NMR) spectroscopy. NMR detects atomic nuclei with nonzero spin; the signals produced by these nuclei are shifted in the magnetic field depending on their electronic environment. By interpreting the chemical shifts observed in the NMR spectrum of a molecule, distances between particular atoms in the molecule can be estimated. To be studied using NMR, a protein needs to be small enough to rotate rapidly in solution (on the order of 30 kilodaltons in molecular weight),

soluble at high concentrations, and stable at room temperature for time periods as long as several days.

Analysis of the chemical shift data from an NMR experiment produces a set of distance constraints between labeled atoms in a protein. Solving an NMR structure means producing a model or set of models that manage to satisfy all the known distance constraints determined by the NMR experiment, as well as the general stereochemical constraints that govern protein structures. NMR models are often released in groups of 20 -40 models, because the solution to a structure determined by NMR is more ambiguous than the solution to a structure determined by crystallography. An NMR average structure is created by averaging such a group of Models. Depending on how this averaging is done, stereochemical errors may be introduced into the resulting structure, so it's generally wise to check the quality of average structures before using them in modeling.

Check your progress-3

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. Explain the structure determination methods

………………………………………………………………………………………………

1.8 Feature Detection

Features in protein sequences represent the details of the protein's function. These usually include sites of post-translational modifications and localization signals. Post- translational modifications are chemical changes made to the protein after it's transcribed from a messenger

RNA. They include truncations of the protein (cleavages) and the addition of a chemical group to regulate the behavior of the protein (phosphorylation, glycosylation, and acetylation are common examples).

Localization or targeting signals are used by the cell to ensure that proteins are in the right place at the right time. They include nuclear localization signals, ER targeting peptides, and transmembrane helices. Protein sequence feature detection is often the first problem in computational biology tackled by molecular biologists. Unfortunately, the software tools used for feature detection aren't centrally located. They are scattered across the Web on personal or laboratory sites, and to find them you'll need to do some digging using your favorite search engine.

One collection of web-based resources for protein sequence feature detection is maintained at the Technical University of Denmark's Center for Biological Sequence Analysis Prediction Servers (http://www.cbs.dtu.dk/services/). This site provides access to server-based programs for finding (among other things) cleavage sites, glycosylation sites, and subcellular localization signals. These tools all work similarly; they search for simple sequence patterns, or profiles, that are known to be associated with various post-translational modifications. The programs have standardized interfaces and are straightforward to use: each has a submission form into which you paste your sequence of interest, and each returns an HTML page containing the results.

1.9 Secondary Structure Prediction

Secondary structure prediction is often regarded as a first step in predicting the structure of a protein. Secondary structure prediction methods can be divided into alignment-based and single sequence-based methods. Alignment-based secondary structure prediction is quite intuitive and related to the concept of sequence profiles. In an alignment-based secondary structure prediction,

the investigator finds a family of sequences that are similar to the unknown. The homologous regions in the family of sequences are then assumed to share the same secondary structure and the prediction is made not based on one sequence but on the consensus of all the sequences in the set. Single sequence-based approaches, on the other hand, predict local structure for only one unknown.

Modern methods for secondary structure prediction exploit information from multiple sequence alignments, or combinations of predictions from multiple methods, or both. These methods claim accuracy in the 70 -77% range. Many of these programs are available for use over the Web. They take a sequence (usually in FASTA format) as input, execute some procedure on them, then return a prediction, usually by email, since the prediction algorithms are compute-intensive and tend to be run in a queue. The following is a list of most commonly used methods:

1.9.1 PHD

PHD combines results from a number of neural networks, each of which predicts the secondary structure of a residue based on the local sequence context and on global sequence characteristics (protein length, amino acid frequencies, etc.) The final prediction, then, is an arithmetic average of the output of each of these neural networks. Such combination schemes are known as jury decision or (more colloquially) winner-take-all methods. PHD is regarded as a standard for secondary structure prediction.

1.9.2 PSIPRED

PSIPRED combines neural network predictions with a multiple sequence alignment derived from a PSI-BLAST database search. PSIPRED was one of the top performers on secondary structure prediction at CASP 3.

2.0 Predicting 3D strcture and protein modeling

The main goal of homology modeling is to predict a structure from its sequence based on the similarity and identity. The target sequence will be retrieved from the NCBI. Further, the target sequences will be compared with known structures stored in the PDB (using BLAST). In case if any protein structures with 50% identical residues with the target sequence, then the sequences will be aligned and finally we arrive at model structure. In practice, homology modeling is a multistep process that can be divided into following steps:

1. Template identification and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side chain modeling
6. Model optimization
7. Model validation

2.0.1 Template identification and initial alignment

The percentage of identity between the target sequence and a possible template should be high enough with simple sequence alignment programs such as BLAST (Altschul *et al*., 1990) or FASTA (Pearson, 1990). Two different matrices namely residue exchange matrix and alignment matrix can be used for this function.

2.0.2 Alignment Correction

In some cases, it is very difficult to align two sequences in a region where the percentage of sequence identity is very low. In such cases many programs (e.g. CLUSTALW) can be used to align the sequences. It provides some additional information on sequence alignment. Multiple

sequence alignments are outstanding technique in homology modeling, for example, to place deletions (missing residues in the model) or insertions (additional residues in the model) in sequences are highly divergent region.

2.0.3 Backbone Generation

After aligning the sequence, we can initiate the actual model building. Two different procedures are available: 1) The backbone coordinates (N, C, Cα and O) only be copied if two aligned residue differ, 2) the two aligned residues are same then side chains also can be included.

2.0.4 Loop Modeling

The target and template alignment structure contains gaps in majority of cases. These gaps may come from either in the model sequence or in the template sequence. In such case, there two different methods such as knowledge based and energy based methods are used for the loop modeling.

Knowledge based: It searches the PDB for known loops with endpoints that match the residue between with the loop has to be inserted, and simply copies the loop information (Eg: 3D-Jigsaw, Insight, Modeller, Swiss-model).

Energy based: In this case an energy function has used to judge the quality of loop by *ab inito* fold prediction. Further, the energy was minimized using Monte Carlo or molecular dynamics simulation techniques to get the best loop conformation.

2.0.5 Side-Chain Modeling

The side-chain conformations (rotamers) of residues are conserved in structurally similar proteins. They often have similar χ_1 – angles (i.e., the torsion angle about the C_α - C_β bond). Thus, it is possible to copy the conserved residues entirely from the template to the model. It

produces higher accuracy structure than by copying just the backbone and repredicting the side chains.

2.0.6 Model Optimization

The backbone correction is the critical step in model building since their turn depends on the rotamers and their packing. Model optimization is a significant procedure since there are more possibilities to mislead the co-ordinates of the generated structure. Two different ways to achieve the precision:

1. Qunatum force fields: Protein force fields should be fast to handle large molecules efficiently. Thus energies are normally expressed as function of the atomic nuclei positions. More accurate electrostatics has applied to derive the classical force fields.

2. Self-paramaterizing force fields: The large extent of force field parameters affects the force field accuracy. These force fields are usually attained from quantum chemical calculations (e.g., Van der Waals radii, atomic charges).

However, the most straightforward approach to model optimization is to run a MD simulation of the model. This process mimics true folding process. Thus, the model will complete its folding and "home in" to the true structure during the simulation.

Model validation

In most of the cases the generated models may contain errors. The number of errors mainly depends on two values:

1. The percentage sequence identity between target and template. If the similarity is > 90% than the accuracy of the model can be compared to crystallographically determined structures. In case of 50% - 90% identity, the rms error in the modeled co-ordinated can be > 1.5 Å. If the sequence identity is < 25%, the alignment turns out to be the main bottleneck for homology modeling.

There are two different ways to estimate errors in a structure:

1. Calculating the models energy based on a force field: This methods check the bond length and bond angle are within the normal ranges.

2. Determination of normality of the modeled structure on how well a given characteristic resembles the same characteristic in real structures. Most of them are mainly depend on the interatomic distances and contacts.

Check your progress-4

Note: a. write your answer in the space given below

b. Compare your answer with those given at the end of the unit

i. List out the various steps of protein modelling

…………………………………………………………………………………………

2.1 Let Us Sum Up

In this unit you have learnt about the structure prediction, determination, and pathway analysis, protein information and modeling. This knowledge might a better understanding about the protein structure. The databases and tools would be helpful to identify the particular protein and their sequence, structure and functional annotation. This information is also used to predict the structure of a protein.

2.2 Answers to check your progress

1. Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that help students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease. SWISS-PROT

is a protein knowledge based database that contains amino acid sequence, protein name, taxonomic data, and citation information.

2. WIT, PathDB and KEGG
3. X-ray Crystallography and NMR Spectroscopy
4. Template identification and initial alignment, Alignment correction, Backbone generation, Loop modeling, Side chain modeling, Model optimization, Model validation

<u>Distance Education-CBCS-(2019-20 Academic Year Onwards)</u>

M.Sc Zoology Question Paper Pattern- Theory

Bioinformatics & Biostatistics (Course Code: 36443)

Time: 3 Hours Maximum: 75 Marks

Part – A (10 x 2 = 20 Marks)

Answer all questions

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (a) or (b)

11. a.

(or)

b.

12. a.

(or)

b.

13. a.

(or)

b.

14. a.

(or)

b.

15. a.

(or)

b.

Part – C (3 x 10 = 30 Marks)

Answer all questions

16.

17.

18.

19.

20.